Oncogenomic Pattern Detection in Cancer Copy Number Alteration Data for Pathway Description and Disease Classification

Dissertation zur Erlangung der
Naturwissenschaftlichen Doktorwürde
Dr. sc. nat.

vorgelegt der
Mathematisch-naturewissenschaftlichen Fakultät
der
Universität Zürich

von
Nitin Kumar
aus Indien

Promotionskomitee
Dr. Michael Baudis
(Leiter der Dissertation)
Prof. Dr. Niko Beerenwinkel
Prof. Dr. Josef Jiricny
Prof. Dr. Christian von Mering
Dr. Hubert Rehauer
Prof. Dr. Andreas Wagner
(Vorsitz der Dissertation)

Zürich 2012

Contents

1 Summary

Cancer is a condition associated with changes in the genome of the affected cells. These genomic changes can be somatic (present only in cancer cells) or inherited (present in all cells). The landscape of these changes can vary from single base mutations to copy number alterations (CNA) to changes of the genome structure. Besides changes in the genetic code, cancer cells also acquire epigenetic changes. Analysis of cancer genomes can give detailed insights into the mechanisms of cancer development and progression.

CNA are one type of genomic change reported in nearly all neoplasias. Most of the analyses involving CNA data have been cancer type specific and dealt with identification of frequently copy number altered regions. CNA profiles across samples have also been used to find heterogeneous groups in tumor sample data. The research community is in need of tools for the analysis of CNA data from several systems biology aspects. In this thesis, I present various methods I have used for analysis of CNA data.

The first tool developed by me involved finding co-occurring CNA across cancer samples. The analysis resulted in the formulation of an algorithm, CDCOCA (complexity dependence of co-occurring chromosomal aberrations) which accounts for genomic sample heterogeneity during the determination of statistical relevance of co-occurring CNA. The major strength of the method lies in the detection of relatively infrequent, but statistically highly correlated CNA events. These CNA may be overlooked in conventional analyses, but potentially point towards pathophysiological mechanisms relevant for cancer development. While looking for gene associated signals in enriched genomic regions it was observed that multiple genes for cancer associated pathways were present in these co-occurring CNA.

Signaling pathways are known to be altered by genomic events during neoplastic transformation. One of my projects explored how CNA target signaling pathways. Two new methods for pathway enrichment analysis of CNA data, “G-path” and “S-path”, were developed. As a unique feature of the method, S-path considers the issue of the genomic clustering of genes from cellular pathways while deriving the statistical significance of pathway enrichment. Both new

methods outperformed traditional techniques used for pathway enrichment. Pathway signatures altered in a large heterogeneous dataset as well as across several cancer types were identified using S-path.

Whole genome CNA patterns differ between distinct cancer entities. Some evidence points towards recurring and possibly tissue type related CNA elements. As a sub-project of my work I explored the relative specificity of CNA elements for different tumor entities. The methodology is based on the hierarchical clustering of tumor type specific CNA frequency profiles with a region specific signal randomization. Measuring region dependent perturbations of the initial dendrogram statistics are used to identify "non-neutral" regions, whose propensity for copy number alterations varied with the type of cancer at hand. In an analysis of data from 160 cancer entities, only a subset of these non-neutral loci overlapped with earlier implied, highly recurrent ("hot-spot") cytogenetic imbalance regions.

The work presented in this thesis gives important directions to a systemic analysis of CNA data from different aspects, e.g. complexity dependence of co-occurring CNA, pathway alterations by CNA and the determination of CNA driving cancer to cancer divergence. The statistical tools presented here should prove instrumental in the statistically sound analysis of large scale cancer genome data, with implications for understanding the complexity of human malignancies.

2 Zusammenfassung

Krebserkrankungen gehen mit Veränderungen im Genom der betroffenen Zellen einher. Diese Veränderungen können entweder somatisch (d.h. nur in den Krebszellen vorhanden) oder aber vererbt (d.h. in allen Zellen des körpers) sein. Die Landschaft dieser Veränderungen variiert von Mutationen einzelner DNS-Basen über regionale Veränderungen der Anzahl der DNS-Kopien ("copy number aberrations", CNA) hin zu strukturellen Veränderungen des Tumor zellgenoms. Die Analyse von Krebsgenomen ermöglicht den Einblick in die Mechanismen von Krebsentstehung und -progression.

CNA sind eine Art von Veränderungen der Tumorgenome, welche in der Mehrzahl der Neoplasien festgestellt werden konnte. Die Mehrzahl der bisher berichteten CNA-Analysen war spezifisch für einzelne Entitäten und fokussiert auf die Bestimmung häufig betroffener Regionen. Der Vergleich von CNA-Profilen multipler individueller Neoplasien wurde bereits zur Einschätzung der genomischen Heterogenität innerhalb kliniko-pathologischer Entitäten verwendet. Bisher fehlen allerdings Werkzeuge zur Analyse komplexer CNA-Daten aus einer systembiologischen Perspektive. In dieser Arbeit präsentiere ich verschiedene Methoden die ich zur Analyse von CNA Daten evaluiert oder selbst entwickelt habe.

Die erste von mir entwickelte Methode war auf die Bestimmung gemeinsam auftretender CNA gerichtet. Dieses Teilprojekt resultierte in der Formulierung eines Algorithmus "CDCOCA" (complexity dependence of co-occurring chromosomal aberrations; Komplexizitätsabhängigkeit gemeinsam auftretender chromosomaler Aberrationen). Dieses biostatistische Verfahren bezieht erstmalig die genomische Heterogenität der einzelnen Tumoren als wichtigen Parameter zur Bestimmung statistisch relevanter Kombinationen individueller CNA ein. Die Stärke dieses Verfahrens liegt in der Bestimmung relativ selten auftretender, aber statistisch hochkorrelierter CNA-Ereignisse, welche auf für die Neoplasieentstehung relevante pathophysiologische Methanismen hinweisen können. In der Tat konnten in mittels CDCOCA identifizierten CNA tumorassoziierte Gene gefunden werden.

Zelluläre Signalkaskaden sind ein bekanntes Ziel onkogenomischer Mutationen. Ein weit-

eres Teilprojekt meiner Arbeit befasste sich mit der Entwicklung biostatistischer Methoden zur Bestimmung von häufig von CNA betroffenen Signalkaskaden, wobei bei der Methodenentwicklung die potentielle genomische Gruppierung funktionell verwandter Gene berücksichtigt wurde. In der Anwendung zeigten sich die beiden Methoden "G-Path" und "S-Path" gegenüber traditionellen Verfahren zur Bestimmung relevanter Signalkaskaden überlegen. Sowohl in einem grossen, heterogenen Datensatz als auch in einzelnen Entitäten konnte S-Path Signaturen von Signalkaskaden identifizieren.

Gesamtgenomische CNA-Muster differieren zwischen unterschiedlichen Tumorentitäten, wobei sich Hinweise auf wiederkehrende und gewebsassoziierte Elemente finden lassen. In einem weiteren Teilprojekt untersuchte ich die relative Spezifität von CNA-Elementen für Tumorentitäten. Das dazu entwickelte Verfahren basiert auf einer hierarchischen Clusteranalyse regionaler CNA-Frequenzen in multiplen Tumorentitäten, gefolgt von einer systematischen Randomisierung der regionalen Werte. Die Perturbationen der initialen Dendrogrammstatistik dienen hierbei als Indikator der Spezifität der regionalen CNA für einzelne oder multiple Tumorentitäten ("nicht neutral"). Die auf einer Sammlung der CNA-Profile von 160 Tumorentitäten basierenden "nicht neutralen" CNA entsprachen dabei nur teilweise vorbeschriebenen hochrekurrenten CNA Ereignissen und können auf tumortypspezifische Mechanismen hindeuten.

Die hier präsentierte Arbeit bietet wichtige Beiträge zur systematischer Analyse verschiedener Aspekte von CNA-Daten, zum Beispiel der Feststellung gekoppelter CNA-Ereignisse, von durch CNA präferentiell veränderten Signalkaskaden als auch der Bestimmung von tumortypspezifischen regionalen genomischen Ereignissen. Die neu entwinckelten statistischen Werkzeuge sollten sich als relevant für die systematische Analyse grosser Tumordatensätze erweisen, und Bedeutung für das Verständnis der Komplexität maligner Neoplasien erlangen.

3 Introduction

3.1 Cancer as a disease

Neoplasias are a class of disease marked by uncontrolled and abnormal cell proliferation. Neoplasia may from localized masses of cells, termed "tumors". A neoplasm with a confined growth is typically referred to as "benign". Some neoplastic cells can become invasive and start infiltrating the surrounding normal tissues. This condition is termed as malignancy. Neoplastic cells can also invade the lymphatic and/or the blood system and migrate to organs such as lymph nodes, liver, brain, lungs and bones, leading to formation of tumors in these organs. The spreading tumors are termed "metastiatic". Lymph nodes can either act as mode of distant transmission or a secondary organ in which cancer cells can grow.

Cancer is the second most common cause of deaths following cardiovascular diseases [1], causing one in eight deaths worldwide [2]. Cancer can develop in the majority of human cell types and tissues (Figure 1). Cancer classification systems relate to the type of tissue type in which a cancer originates (histology) or the primary site (body location in which cancer first develops). For ease of understanding, cancer can be separated into a number of major categories: [3]

1. **Carcinoma** - Carcinomas are the cancers of epithelial origin which consists of the inner and outer lining of the human body(e.g. skin and lining of organs), accounting for about 80-90% of human cancers. Examples include breast, prostrate and lung carcinomas. Carcinomas can broadly be sub-divided as - a) squamous cell carcinoma - originating in squamous epithelium and b) adenocarcinoma - developing in an organ or a gland.

2. **Sarcoma** - Sarcomas are the cancers of supportive and connective tissue originating from the embryonic mesoderm. Examples of sarcomas are Osteosarcoma (bone tumors) and Liposarcoma (tumors of the adipose tissue).

3. **Leukemia, Lymphoma and Myeloma** - This category consists of malignancies originating in the hematopoietic system. Leukemias develop in the bone marrow and is charac-

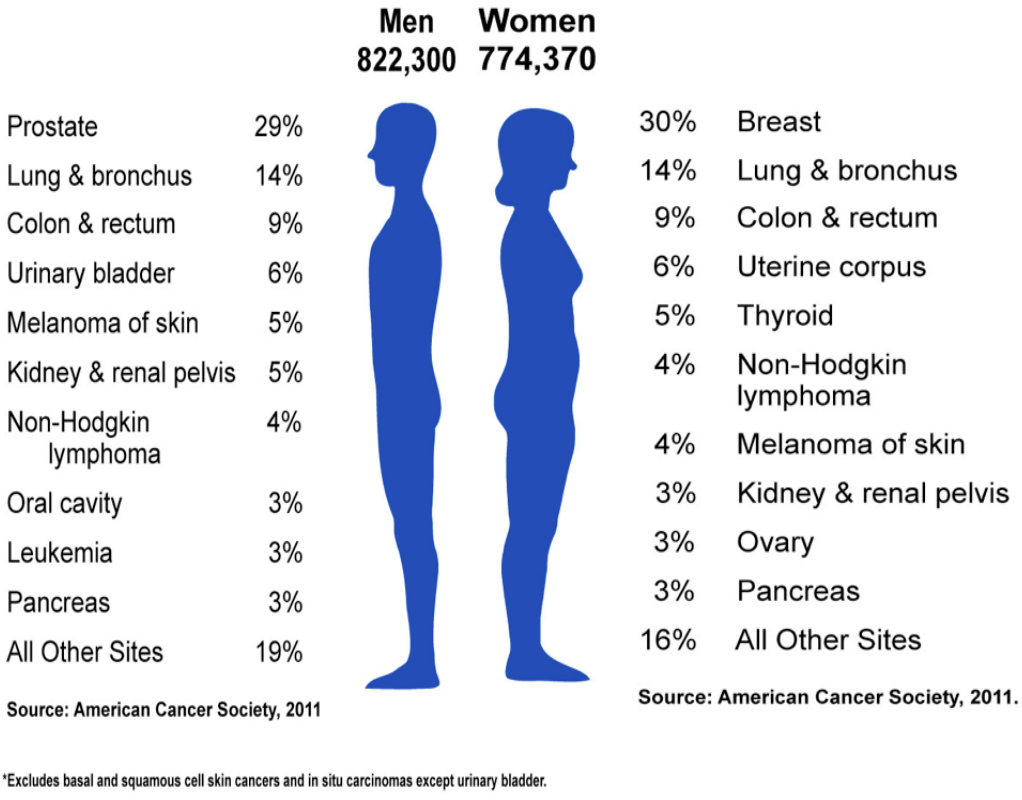


Figure 1: Estimated US cancer cases 2011 [4]. Relative frequency of cancer occurrence among different sites in human body between men and women.

terized by abnormal production of anomalous white blood cells. When leukemia develops from an immature blood cell it is termed as "acute leukemia" (e.g. Acute lymphoblastic leukemia (ALL)) whereas leukemia developing from mature blood cells is termed as "chronic leukemia" (e.g. Chronic myelogenous leukemia).

Lymphoma is a cancer of lymphocytes and develops in lymph nodes, causing them to expand. Malignant lymphomas consist of a large number of different entities, which are defined based on clinical, phenotypic, and molecular characteristics according to the World Health Organization (Table 1). The major groups are Mature B cell neoplasms (e.g.

Diffuse large B cell lymphoma), Mature T cell, and natural killer (NK) cell neoplasms (e.g. Adult T cell leukemia/lymphoma) and Hodgkin lymphoma (e.g. Nodular lymphocyte-predominant Hodgkin lymphoma).

Myeloma is a cancer of the plasma cells, a type of white blood cells producing antibodies e.g. multiple myeloma.

4. **Germ cell tumor** - Germ cell tumors consist of cells matching one or more of the 3 germ layers. Germ layer tumors also called as teratoma, can also consist of normal tissues found in lung, liver and brain.

Table 1: **Examples of malignant lymphomas**

Mature B cell neoplasms	**Mature T cell and natural killer (NK) cell neoplasms**	**Hodgkin lymphoma**
Chronic lymphocytic leukemia /small lymphocytic lymphoma	T-cell prolymphocytic leukemia	Nodular lymphocyte-predominant Hodgkin lymphoma
Splenic marginal zone lymphoma	T-cell large granular lymphocytic leukemia	Classical Hodgkin lymphoma
Hairy cell leukemia	Aggressive NK cell leukemia	
Plasma cell myeloma	Subcutaneous panniculitis-like T-cell lymphoma	
Splenic lymphoma/leukemia, unclassifiable	Adult T-cell leukemia/ lymphoma	
Extranodal marginal zone B-cell lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma)	Sézary syndrome	
Nodal marginal zone B-cell lymphoma (MZL)	Primary cutaneous gamma-delta T-cell lymphoma	
Follicular lymphoma	Anaplastic large cell lymphoma (ALCL), ALK	
Mantle cell lymphoma	Systemic EBV+ T-cell lymphoproliferative disease of childhood	
Diffuse large B-cell lymphoma (DLBCL), not otherwise specified	Hepatosplenic T-cell lymphoma	
Burkitt lymphoma	Mycosis fungoides	

Cancer develops due to mutations in genome affecting different kinds of genes. The rate

of mutation in cells can be enhanced by the presence of various etiological factors. In the following section the standard cancer etiology i.e. factors which are known to increase the rate of carcinogenesis are discussed.

3.2 Cancer etiology

Cancer is a group of very complex and diverse diseases, differing in their causes of occurrence. During a regular cell cycle the DNA of any given cell may acquire errors (mutations) which are corrected by error correction machinery of the cell. However this machinery is not 100% efficient and still some errors are propagated to the daughter cells. The efficiency of error correction machinery is reduced in several environments (such as exposure to carcinogens), resulting in a higher mutation accumulation rate.

Major external factors associated with cancer development can be grouped in the following categories.

1. **Bacteria associated carcinogenesis** - Bacterial infections are rarely associated with cancer development. A notable exception is the infection with *Helicobacter pylori*. Studies have shown that *Helicobacter pylori* infected individuals have an increased risk of developing gastric adenocarcinomas or mucosa associated B-NHL[5, 6, 7]. *H.pylori* produces a protein CagA which helps the bacteria to attach with the stomach lining. Exposure to CagA causes inflammation, increasing the risk of cancer. It is shown that individuals infected with CagA+ bacteria have a higher risk of developing gastric cancers [5].

2. **Virus associated carcinogenesis** - Around 15% of human cancers are known to be caused by viruses [8]. A virus can ideally induce cancer development either by carrying an overactive oncogene called viral oncogene ("v-onc") into the host cell or by activating a cellular oncogene (proto-oncogene).

   However the majority of virus related human malignancies are based on the inflammation caused by chronic infections. Chronic infection by hepatitis C virus (RNA virus) and hepatitis B virus (DNA virus) can result in cirrhosis, leading to primary hepatocellular

carcinoma. Persistent HPV (Human papilloma virus) infection can cause cervical cancer [9], resulting in deaths of one-third diagnosed cases [10]. HPV has also been associated with oral and anal squamous cell carcinomas. An effective vaccination has shown to prevent infections against HPV [11]. Two vaccines Gardasil and Cervarix have been approved by Food and Drug Administration (FDA) against HPV.

Other virus infections may predispose to cancer through a modulation of the host's immune system e.g. Epstein-Barr virus (EBV) is associated with many malignancies such as B and T cell lymphomas and Hodgkin's lymphoma [12]. HIV (Human immunodeficiency virus) is associated with malignancies e.g. Kaposi's sarcoma and non-Hodgkin's lymphoma. HIV infection compromises the immune system, which can further result in an increased risk of other cancer associated infections e.g. HHV-8, EBV, HBV [13, 14].

3. **Mutagenic agents** - Mutagenic agents increase the genome mutation rate which can lead to cancer development. DNA damaging agents can cause multiple kinds of genomic errors (Table2). Tobacco smoke is associated with several cancers e.g. lung, larynx and head and neck carcinoma. Tobacco contains several carcinogens such as nitrosamines, polycyclic aromatic hydrocarbons and benzene. It has been observed that tobacco can increase the risk of mutations in the human genome by causing DNA damage [15, 16]. Ionizing radiation (e.g X-rays) causes DNA damage either by generation of active radicals (e.g. reactive oxygen specie) or by directly breaking the DNA molecules. Radiations such as UV induce thymine dimerization which can increase the overall genome mutation rate as not all dimers are efficiently corrected by DNA repair enzymes.

3.3 Cancer associated genes

Normal replication processes as well as exposure to mutagenic substances and microbial infections can lead to DNA damage which increases the genome mutation accumulation rate. These mutations can affect any gene in the genome. However, alterations in some cancer associated genes can lead to cancer development (termed as "drivers", contrary to those not causing can-

Table 2: **DNA damaging agents**

Type of DNA damaging agent	**Examples**	**Toxic lesion**
Radiotherapy and radiomimetics	Ionizing radiation, Biomycin	Single strange break, double strand break, base damage
Antimetabolites	5-Fluorouracil, Thiopurines, Folate analogues	Base damage, replication lesions
Replication inhibitors	Aphidicolin, Hydroxyures	Double strand break, replication lesions
Bifunctional alkylators	Nitrogen mustard, Cisplatin	Double strand breaks, DNA cross links, Replication lesions

cer, "passengers"). Cancer development associated genes are broadly classified into two major categories.

3.3.1 Oncogenes

Oncogenes are defined through their cancer promoting activity when activated in the wrong context or without proper regulation. They are either over-expressed in cancer and/or mutations make them constitutively active.

One of the first evidences of oncogene activation came through the study in Burkitt's lymphoma (BL). In BL, the MYC oncogene on chromosome band 8q24 was found to be translocated to immunoglobulin loci on chromosomes 14q, 22q and 2p, leading to an over-expression of MYC. This lead to the conclusion that MYC translocation could be an initiating event [17]. Another oncogene commonly activated by translocation is BCL2 in follicular and diffuse large cell lymphomas [18, 19]. In comparison to MYC and BCL2, oncogenic activity of KRAS oncogene was identified using transfection experiments. Mouse fibroblasts were transfected with DNA from human cancer cells, leading to development of cancer like properties in fibroblast cells. The KRAS gene was found to be responsible for this transformation activity [20, 21]. A few more examples of oncogenes are listed in Tables 3 and 4.

Oncogenes can be activated via several mechanisms such as point mutations, translocations, copy number alterations (CNA) and epigenetic changes. All the mentioned changes lead to either

an over-expression and/or constant activation of an oncogene. Point mutational activation is observed in BRAF, where the mutation converts a valine present in kinase domain of BRAF to glutamate and alters the phosphorylation of adjacent residues, rendering the gene constitutively active [22, 23, 24]. BRAF also attains an over-expression by gain CNA in astrocytomas [25, 26].

Oncogenes can have a variety of functions in a cell. Some genes such as TMPR552 function as a transcription factors. In prostrate cancer it often fuses with ETV1 making it constitutively active, which further activates or represses genes involved in differentiation and apoptosis [27]. Other examples of transcription factor genes include MYC, REL and ERBA1. Oncogenes such as platelet derived growth factor (PDGF) function as growth factors and their over-expression can induce transformation in cells over-expressing PDGFR receptors [28]. Oncogenes can also function as growth factor receptors (e.g. Epidermal growth factor receptor (EGFR)) and signal transducers e.g. (ABL, SRC, PI3K) [29].

Oncogene derived proteins can potentially act as drug targets for cancer treatment, as cancer cells may depend on that protein's production for their survival. For example, oncogenic proteins can be targeted by molecular markers or antibodies (cell surface proteins) to hinder the growth of cancer cells. Examples are Imanitib which affects KIT and PDGFR receptor kinases [30], gasatinib, an inhibitor of SRC tyrosine kinase family [31] and gefitinib which is an EGFR inhibitor [32].

3.3.2 Tumor suppressor genes

Unlike oncogenes, alterations in tumor suppressor genes are aimed at reducing their activity. This reduction can come either via a deactivating mutation or loss of gene copies affecting the protein amount. The idea about existence of tumor suppressor genes came from a somatic fusion experiment. Fusion of cancer cells with normal cells reverted cancer cells to normal phenotype [33]. Further studies on retinoblastoma tumors in children led to the "two hit hypothesis" model of tumor suppression [34] (Figure 2) in which both the alleles of a gene should be mutated to produce a cancer phenotype. In hereditary retinoblastoma one mutated allele of RB1 gene is inherited and the other is somatically mutated leading to cancer development. The two hit model

Table 3: **Germ line mutated cancer genes and associated syndromes**

Gene	**Associated Syndrome**	**Gene type**	**Mutation type**
ALK	Familial neuroblastoma	oncogene	T, Mis, A
RET	Multiple endocrine neoplasia 2A/2B	oncogene	T, Mis, N, F
CDK4	Familial malignant melanoma	oncogene	Mis
KIT	Familial gastrointestinal stromal tumor	oncogene	Mis, O
APC	Adenomatous polyposis coli; Turcot syndrome	tumor-suppressor	D, Mis, N, F, S
BRCA1	Hereditary breast/ovarian cancer	tumor-suppressor	D, Mis, N, F, S
BRCA2	Hereditary breast/ovarian cancer	tumor-suppressor	D, Mis, N, F, S
BLM	Bloom Syndrome	tumor-suppressor	Mis, N, F
EGFR	Familial lung cancer	oncogene	A, O, Mis
FANCD2	Fanconi anaemia D2	tumor-suppressor	D, Mis, N, F
MLH1	Hereditary non-polyposis colorectal cancer, Turcot syndrome	tumor-suppressor	D, Mis, N, F, S
MSH2	Hereditary non-polyposis colorectal cancer	tumor-suppressor	D, Mis, N, F, S
RB1	Familial retinoblastoma	tumor-suppressor	D, Mis, N, F, S
TP53	Li-Fraumeni syndrome	tumor-suppressor	D,Mis, N, F

Examples of cancer associated genes affected by hereditary mutations, the name of associated syndromes and the mutation type they are often altered with.

was validated by identification of a deletion locus involving chromosome 13 in retinoblastoma. The gene RB1 present in this locus was subsequently cloned [35, 36, 37]. Due to the two hit model most of the early attempts were made to identify biallelically deactivated tumor suppressor genes e.g. APC [38], BRCA1[39] and BRCA2 [40].

In recent years it has been also established that not all tumor suppressor genes follow a two hit model and some show "haploinsuffiency", i.e. inactivation of a single copy of tumor suppressor gene is enough to facilitate cancer development. Examples for genes with relevant haploinsuffiency are e.g. BRCA1, BRCA2 [41, 42, 43] and PTEN[44] (Figure 2).

Like oncogenes, tumor suppressor genes can also have a variety of functions inside a cell.

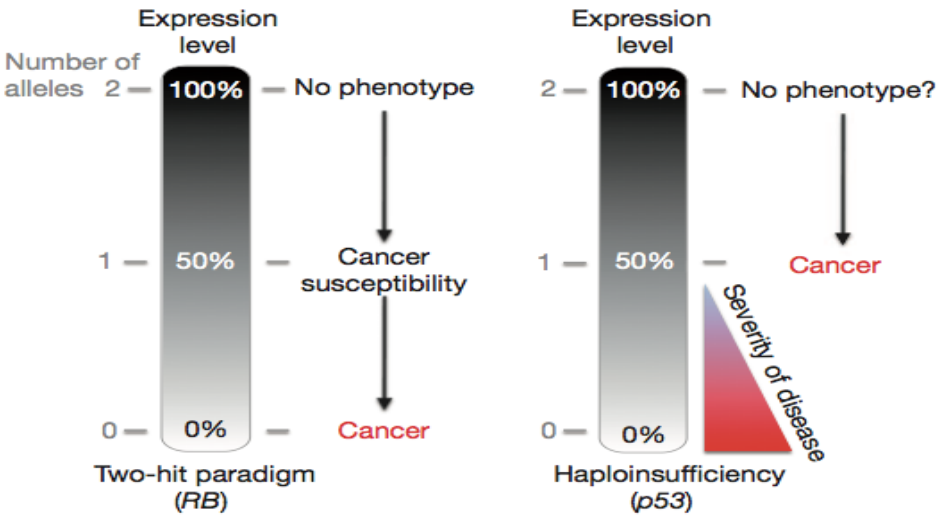


Figure 2: Tumor suppression model. RB follows a two-hit paradigm where loss of both ales induces cancer. This loss can be germ line or somatic. TP53 whereas shows haploinsuffiency, where loss of one allele is sufficient to induce cancer formation [44].

Some tumor suppresor genes such as RB1 act as transcriptional corepressors. RB1 interacts with transcription factors and regulates expression of genes required for replication and DNA metabolism. A loss of RB1 leads to disassociation of external signal to cell cycle machinery, leading to cell proliferation [45]. Others such as TP53 are considered as guardians of the genome. It is a transcription factor, involved in processes e.g. DNA repair and apoptosis [46]. Others e.g. APC and SMAD4 act as signal transduction molecules.

Recently another class of tumor suppressor genes, "stability/caretakers genes" has been defined. Examples include genes involved in mismatch repair, base-excision repair, and nucleotide repair pathways e.g. BRCA1 and ATM affect processes involving chromosomal segregation and recombination, hence affecting overall genomic stability. When stability genes are affected, overall mutation accumulation rate of a cell increases, leading to mutations in oncogenes and tumor suppressor genes eventually causing cancer. Germline mutations in stability genes e.g. BLM (associated with Bloom syndrome) [47, 48] and BRCA1/2 [49, 50] predispose humans to

several cancers (Table 3).

Most of the mutations in cancer happen somatically i.e. only the cancer tissue shows their presence. However germline mutations (mutations in germ cells and inherited) in cancer genes predispose different human tissues to cancers. Some hereditary gene mutations increase the risk of multiple cancers in different parts of human body (Table 3). People with hereditary mutations in disease causing genes are categorized as "genetic susceptible" as they have a higher rise of developing cancer than those with no hereditary mutations.

Table 4: **Most common somatically altered cancer associated genes**

Gene	**Tumour Type**	**Gene type**	**Mutation type**
ABL1	CML, ALL, T-ALL	oncogene	T, Mis
AKT2	ovarian, pancreatic	oncogene	A
ALK	ALCL, NSCLC, Neuroblastoma	oncogene	T, Mis, A
ERBB2	breast, ovarian, other tumor types, NSCLC, gastric	oncogene	A, Mis, O
LMO1	T-ALL, neuroblastoma	oncogene	A, T, A
MDM4	GBM, bladder, retinoblastoma	oncogene	A
MYC	Burkitt lymphoma, amplified in other cancers, B-CLL	oncogene	A, T
MYCN	neuroblastoma	oncogene	A
REL	Hodgkin Lymphoma	oncogene	A
ATM	T-PLL	tumor-suppressor	D, Mis, N, F, S
RB1	retinoblastoma, sarcoma, breast, small cell lung	tumor-suppressor	D, Mis, N, F, S
TP53	breast, colorectal, lung, sarcoma, adrenocortical, glioma, multiple other tumor types	tumor-suppressor	D,Mis, N, F
BAP1	uveal melanoma, breast, NSCLC	tumor-suppressor	N, Mis, F, S, O
CDKN2C	glioma, MM	tumor-suppressor	D
DNMT3A	AML	tumor-suppressor	Mis, F, N, S
NF1	neurofibroma, glioma	tumor-suppressor	D, Mis, N, F, S, O
PIK3R1	gliobastoma, ovarian, colorectal	tumor-suppressor	Mis, F, O
IKZF1	ALL	tumor-suppressor	D
TNFAIP3	marginal zone B-cell lymphomas, Hodgkin's lymphoma, primary mediastinal B cell lymphoma	tumor-suppressor	D, N, F

Examples of cancer associated genes affected by somatic mutations, the cancer type they are commonly mutated in and the mutation type they are often altered with.

Various types of genomic mutations can target multiple genes, resulting in development of different cancer types. This association can either be gene-mutation and/or gene-cancer type specific (Table 4). Cancer genes can be targeted by several genomic changes, altering their expression/activity. The kind of genomic alterations observed in cancer are described in next section.

3.4 Genomic alterations in cancer

Cancer develops due to changes at the genome level. All cells in human body originate from a single fertilized egg through a series of mitotic divisions. Cell division involves replication of cell DNA to form two daughter cells. However, this replication efficiency is not 100% accurate and during the division process cells acquire mutations. Mutations can also happen during the non-replicative phase of cell cycle due to DNA damage, which can happen because of exogenous agents (e.g. tobacco smoke and radiation). Mutations in cancer cell can be "germline" i.e. hereditary and present in all cells of an organism or "somatic", present only in cancer cells and their precursors. Alterations in cancer cells can vary from single base changes (e.g. point mutations) to structural alterations (e.g. translocations) to chromosome arm level changes (e.g. copy number alterations). All these changes can be associated with cancer development and progression [2, 51, 52, 53, 54].

It has been well established that cancer follows an evolutionary dynamics [2] and mutations accumulation in a stepwise manner (Figure 3). Mutation accumulate randomly in the genome followed by selection to produce a phenotype. Cancer cells are selected for a proliferative phenotype either by faster cell divisions, a reduction in cell death and apoptosis or apoptotic/senescence escape. During the process of mutation accumulation (during cell division or repair of damaged DNA) cells acquire changes. Most of these changes are neutral and are removed from the system. However some of the changes can lead to cancer development. The changes associated with cancer development are called "driver" changes and most of these changes happen in oncogenes and/or tumor suppressor genes. Mutations not related to cancer development are termed "passengers" (Figure 3). A very simple way to find drivers is by looking

at their frequency of occurrence at any genomic locus across several samples. Another way is by comparing different stages of cancer development and progression. Mutations which occur more frequent than random could be important for cancer development and can be cancer type specific. Through comparing different tumor stages for somatic alterations a classic tumor progression model has been derived [55, 56]. The cancer evolution paradigm was recently extended by observing massive genomic rearrangements that can accumulate at once in cells [57] and may result in the development of cancer. However, the exact cause of these high number of rearrangements is still questionable.

An individual cancer cell can acquire all kinds of genomic changes [2, 51, 52, 53, 54, 58]. In the following section we describe the mutational changes observed in cancer.

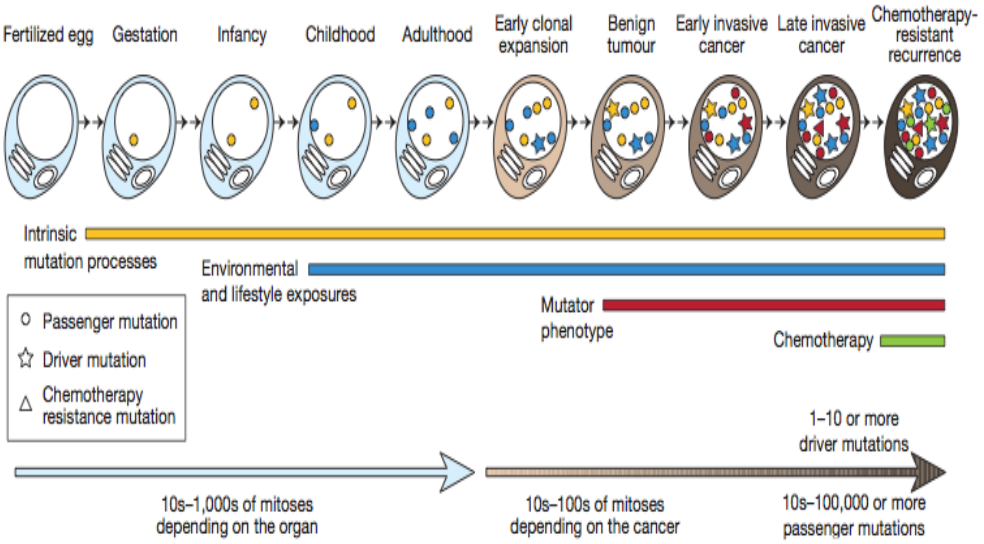


Figure 3: Pattern of mutation accumulation across cells. Some of them called drivers give selective edge to cells and lead to development. Others termed passengers just accumulate during normal cell divisions and has either no role in cancer development or give selective edge to it [2].

3.4.1 Point mutations

Point mutations are a single base substitution, addition or deletion events. They have been described in nearly all cancers studied. Their relative abundance is still a subject of discussion. Consortiums such as TCGA and ICGC have recently sequenced some cancer patient's genomes (as well as cell lines) to access the total amount of point mutations in cancer. However, the studies still need a large amount of genomes to be sequenced to define a point mutation landscape. A base substitution event at any particular exon position in a gene can be synonymous (no change in coded amino acid) or non-synonymous (coded amino acid changes). Non-synonymous mutations can alter the protein activity, increasing the activity in case of oncogenes or lowering/extinguishing for tumor suppressor genes. Most point mutations in a cell are either neutral or detrimental and only very few mutations are beneficial for the development and undergo positive selection.

By affecting a coding amino acid, point mutations can alter the protein folding, stability, function and protein-protein interactions. Examples of oncogene activating point mutations include mutations in KRAS, HRAS and BRAF [29], whereas tumor-suppressor gene deactivating mutations have been observed in TP53, APC and RB1 [59]. Other examples of point mutational activation/deactivation are listed in Table 3 and 4.

Research has been focussed on predicting the functional impact of point mutations to differentiate between mutations affecting protein activity with neutral ones. Several methods for assessing the mutational effect of a change have been developed. These methods compare the properties e.g. size, polarity, structural information of protein and evolutionary conservation of the original and substituted amino acids [60].

3.4.2 Chromosomal rearrangements

Chromosomal rearrangements are a type of genomic change observed in the majority of human neoplasias. Chromosomal translocations involve an exchange of genetic material between two or more chromosomes. Chromosomal inversion are another kind of genomic rearrangement here, a genomic region within a chromosome is broken at two points and is rejoined in inverted position

on the same chromosome. Chromosomal rearrangements can lead to an increase in the gene expression of targeted genes by transferring them to an active promotor or vice versa. They can also lead to formation of constantly active hybrid genes or the deactivation of genes through disruption of coding regions. Chromosomal rearrangements can be identified using techniques such as FISH and SKY. Recurrent translocations have mostly been studied in hematologic malignancies.

One of the first consistent chromosomal abnormalities detected was the presence the of Philadelphia (Ph) chromosome in around 90% of chronic myeloid leukemia patients [61]. Initially this was thought to be chromosome 22q deletion; however later on it was identified as a reciprocal translocation involving chromosome 9 and 22 (t (9;22)(q34;q11)) [62]. The reciprocal translocation involving the ABL gene on chromosome 9 and the BCR gene on chromosome 22 [63] leads to the formation of a constantly active tyrosine kinase. This fusion protein activates other cell cycle proteins and enhances cell division.

Another example of a fusion protein is the dimerization between FOS transcription protein and the JUN transcription factor to form AP1 transcription factor resulting in activation of genes controlling cell division [64, 65]. At present more than 200 fusion proteins have already been identified [66].

MYC was the one of the first oncogenes identified as translocated in BL. The MYC locus on 8q24 can translocate to three different loci on chromosome 14q, 22q and 2p, carrying immunoglobulin genes (t(8;14,q24;q32), t(22;8,q11;q24), t(2;8,p12;q24)) [17, 23, 24, 67]. All the translocations lead to an over-activation of MYC either by transferring MYC to an active immunoglobulin locus (e.g. MYC translocated to IgH locus t(8;14)) or vice versa. Other translocations such as t(14;18)(q32;q21), t(11;14)(q13;q32), t(3;14)(q27;q32) and t(11;18)(q21;q21) are very common in B-cell lymphomas. The most common translocation observed in follicular lymphomas and diffuse large B-cell lymphomas involves the anti-apoptotic gene BCL2 (t(14;18)(q32;q21)) and the immunoglobulin heavy chain locus [18, 19].

Some chromosome translocations are highly specific to a single cancer type e.g. t(15;17) observed only in acute promyelocytic leukemia (APL) [68] and t(1;19) found only in B-cell

precursor acute lymphoblastic leukemia (ALL) [69]. Inherited syndromes can predispose individuals to hematologic malignancies e.g. ataxia-telangiectasia patients harboring ATM mutations are prone to develop chromosomal translocations involving T-cell or immunoglobulin antigen-receptor loci [70].

3.4.3 Copy Number Alterations/Aberrations

Copy number alterations/aberrations are one kind of genomic change reported in nearly all cancer types. A copy number alteration is gain or loss of a genomic region and can range from focal events to whole chromosome level changes (aneuploidy). Alterations involving addition or deletion of a single base can have a possible impact due to a quantitative structural or functional change. This is in contrast to a purely qualitative change e.g. frame shift and point mutations. Arm level (or whole chromosome) changes are observed frequently in cancer however it becomes difficult to identify cancer associated genes in these regions. With recent advances in detection of focal CNA some cancer associated genes have been identified [71, 72, 73].

CNA are a key event in cancer development and progression. This thesis deals with the statistical analysis of cancer CNA data from multiple aspects. In the following sections various mechanisms generating CNA, how they can be detected and their importance in cancer development and progression are discussed.

3.5 Copy number alterations and cancer

The nearly omnivorous presence of CNA across cancer types have made them an ideal candidate to identify and study. CNA can accumulate inside a cell due to several mechanisms, which are described in next section.

3.5.1 Mechanisms generating CNA

CNA can have a variable size, ranging from few base additions/deletions to whole chromosome arm level changes. Depending upon the size of copy number alteration various mechanisms can be associated with their occurrence. The major responsible mechanisms are listed here.

1. **Homologous recombination (HR)** - HR is a major pathway to repair double strand breaks and requires homology search of the damaged DNA with undamaged DNA to repair it. It is majorly involved in S and G2 phase of cell cycle [52, 74, 75]. HR is further sub-divided into 4 categories a) non-allelic homologous recombination (NAHR), b) gene conversion, c) break-induced replication (BIR), d) single-strand annealing (SSA). These mechanisms can repair damaged DNA via slightly different mechanisms [52] creating different kinds and size of genomic changes. Deletions can be produced by NAHR, BIR and SSA (small scale deletions) where as duplications can be produced by NAHR, gene conversion (small duplications) and BIR.

2. **Non-homologous end joining (NHEJ)** - NHEJ does not require (or very less) homology search to repair the damaged DNA. NHEJ is a very important DNA repair mechanism as it can work any time during the cell cycle [52, 74, 75]. NHEJ is associated with small deletions and insertions.

3. **Other mechanisms generating small changes** - a) LINE-1 (Interspersed nuclear elements-1) are retrotransposons which can give rise to copy number changes in embryonic stages. Wnt signaling is associated with the activation of LINE-1 elements [76, 77]. b) DNA repeats carrying regions can be extended/deleted due to DNA polymerase slippage and less efficient mismatch repair pathways.

4. **Chromosome level changes** - CNA involving chromosome level changes can happen mainly in the anaphase of cell cycle due to errors in chromosomal segregation and separation. Mutations in "chromosome stability genes" playing a role in chromosome condensation, kinetochore functions, sister-chromatid cohesion and microtubule assembly have shown to be associated with aneuploidy in cancer. Examples of such genes include BRACA1, BRACA2 and MAD2 [78, 79, 80]. Recently a new phenomenon "chromothripsis" involving shattering of a chromosome in to multiple pieces and than rejoining, has been proposed as an initiating event for cancer development [57]. No direct causes for this phenomenon have been identified yet. However its association to mutations in genes such

as TP53 has been proposed [81].

3.5.2 Techniques for detection of CNA

3.5.2.1 Cancer karyotyping Human chromosomes were already being analyzed in nineteenth century by German scholars [82, 83, 84, 85]. However only in 1956 it was shown that humans have 46 chromosomes [86]. Clinical cancer cytogenetics started in 1960 by identification of the Philadelphia chromosome in the blood of chronic myeloid leukemia patients [61].

Cytogenetics was extended with the use of fluorochromes coupled with alkylating agents in the late 1960s, leading to identification of the complete human karyotype [87]. Quinacrine Mustard (QM) was used as a fluorescent agent to bind interphase nuclei. Chromosomes were grouped into different types based on their fluorescence patterns which was determined photoelectrically. Fluorophore labeling proved to be better than other chromosome identification methods e.g. centromere index or autoradiography. Fluorochrome labeling was replaced by Giemsa banding in 1970s [62] (Figure 4). Giemsa stains AT rich regions in the genome as dark and others light, producing bands on chromosomes. Various parameters such as centromere positioning, length of chromosomes, number of chromosomes and length of telomeres can be easily visualized using this staining. These parameters were then used to compare cancer cells with normal cells to determine cancer specific patterns. In the 1980s it became quite obvious that chromosome abnormalities were an essential feature of cancer cells.

The quality of handling techniques improved with the development of promataphase analysis involving synchronized cells. The most common method is to grow cells in presence of Thymidine to block cell cycle in S phase which can be released either by removal of thymidine or addition of deoxycytidine [88]. The blocked cells can than be analyzed using banding techniques to obtain a better resolution of bands.

Development of karyotyping techniques lead to banding analysis of cancer samples. Over the last decades this data has been deposited and collected by repositories such as Mitelman database [89, 90].

3.5.2.2 Fluorescent in situ hybridization (FISH) Advancement in molecular-cytogenetics techniques led to an increase in the resolution at which chromosome could be analyzed. FISH involved sequence specific probes for labeling of genomic DNA in contrast to unspecific chromosome staining. FISH techniques enabled the localization and detection of specific genome regions [91, 92] in interphase or metaphase of cells. They also laid the foundation of current whole resolution array techniques. An increase in resolution became possible primarily due to the development of polymerase chain reaction (PCR) [93]. Since human DNA is very large, fragmented human DNA is often used in designing FISH probes. The fragments were cloned in form of artificial chromosomes (Bacterial artificial chromosomes - BAC) which could then easily be maintained in replicating bacteria. These fragments can be used to generate specific probes.

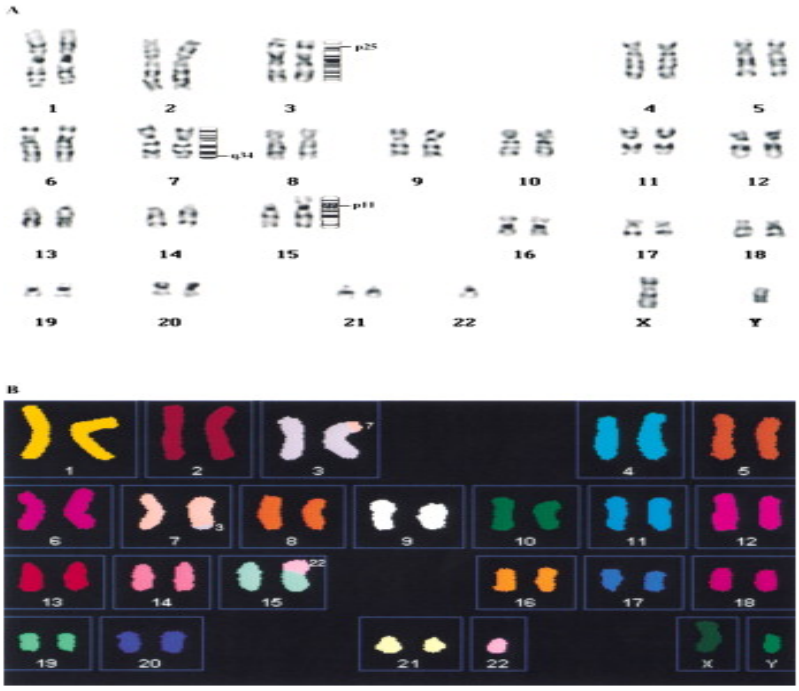


Figure 4: G-banding (A) and SKY (B) karyotype of follicular thyroid carcinoma showing t(3;7)(p25;q34) and dic(15;22)(p11;p11) [94]

The further advancement in FISH techniques came from the increase in number of differ-

entially labelled probes in the same experiment. SKY (spectral karyotyping, Figure 4)[95] and M-FISH (Multiplex Fluorescence In Situ Hybridization) [96] allow for simultaneous visualization of all differentially labelled human chromosomes. Degenerate oligonucleotide-primer PCR (DOP-PCR) is used to produce chromosomes specific painting libraries from flow sorted chromosomes [97, 98] or region specific probes following micro-dissection [99]. Both SKY and M-FISH use same labeling approach, however use different detection technique. In SKY, only one image is captured using epifluorescence microscopy charge-coupled device imaging and Fourier spectroscopy [100], making it possible to measure the entire emission spectrum with a single exposure at all image points. This is in contrast to standard fluorescent imaging where bandpass microscope filters are used to capture separate images for each of the fluorochromes which are than combined.

Genomic abnormalities such as chromosomal breakpoints, translocations and complex rearrangements can be obtained using SKY/M-FISH.

3.5.2.3 Genome wide array techniques In comparison to traditional techniques discussed above, comparative genomic hybridization (CGH [101, 102]) allows for whole genome screening of imbalances. CGH is a molecular-cytogenetic technique for copy number change analysis of a given DNA sample. CGH can detect unbalanced chromosomal changes. CGH is based on the principle of reverse in situ hybridization in which probes are present on a matrix or slide and labelled DNA from tumor and control tissue is hybridized to the probes. Only copy number changes can readily be identified using CGH. In comparison to CGH, array CGH (aCGH) [103, 104] can detect CNA at a higher resolution.

aCGH involves use of bacterial artificial chromosomes (BACs) [105, 106]. The basic principle of aCGH is same as that of cCGH the only difference is that in aCGH metaphase chromosomes have been replaced with relatively high number of mapped clones spotted onto a glass slide. BAC arrays cover entire genome at differently spaced intervals e.g. BCCRC Lam SMRT array 32k_1.2 has nearly 32,000 probes spaced 30Kb apart on genome, CGH-SANGER 3K 3 has around 3K probes 1mb apart each. Analysis using aCGH has led to a reduction in resolution

from about 5-10Mb (for cCGH) to ∼100kb. Instead of large BAC clones DNA sequences (cDNA arrays) [103] can also be used for aCGH. cDNA arrays allow the analysis of both DNA and RNA using same platform however they only covered regions representing exons.

The basic principle of array CGH (aCGH) analysis is illustrated in Figure 5a. Cancer tissue samples and normal genomic DNA are labeled with two different tags (e.g biotin - tumor DNA and digoxigenin - control DNA). Both labelled DNA are mixed with c0t-1 DNA to prevent non-specific hybridization. c0t-1 is placental DNA enriched for repetitive sequences (e.g. Alu and Kpn) of about 50-300bp in length. The mix of three different DNAs is then hybridized to target metaphase chromosomes. The amount of tumor and control DNA bound at any given chromosomal location depends on the the relative abundance of this sequence in both DNA samples. The hybridized DNA is then detected for fluorescence signal using fluorophores (e.g. green-fluorescing fluorescein isothiocynate (FITC) for tumor DNA and red-flourescing rhodamine antidigoxigenin for control DNA), which is than compared to identify abnormal regions. A gain/loss of any genomic region will result in an elevated/reduced green to red ratio respectively. The copy number karyotype of tumor is than generated by comparing DNA from tumor and normal cells.

The resolution of aCGH improved further with the development of oligonucleotide arrays $(60-100\text{bp})$. The first of the arrays were synthesized by mechanically spotting oligonucleotides on to a slide [108] giving a resolution of one probe per 50kb of genome. Recently techniques have been improved and oligonucleotide arrays are directly synthesized on a glass slide. NimbleGen has further improved the resolution of oligonucleotide arrays by spotting around 2.1 million oligonucleotides on the arrays which can detect CNA in range of ∼5kb. The biggest disadvantage of oligonucleotide arrays is poor signal-to-noise ratio. The log2 ratio has a standard deviation of around 0.25 in comparison to BAC arrays which is around 0.05.

An advancement in the analysis of CNA came from development of genotyping arrays which could screen for several million probes simultaneously with the help of single nucleotide polymorphisms [109, 110, 111, 112] arrays. The arrays were primarily designed for genotyping SNPs, however they have proven to be very useful for copy number detection. Unlike aCGH, SNP

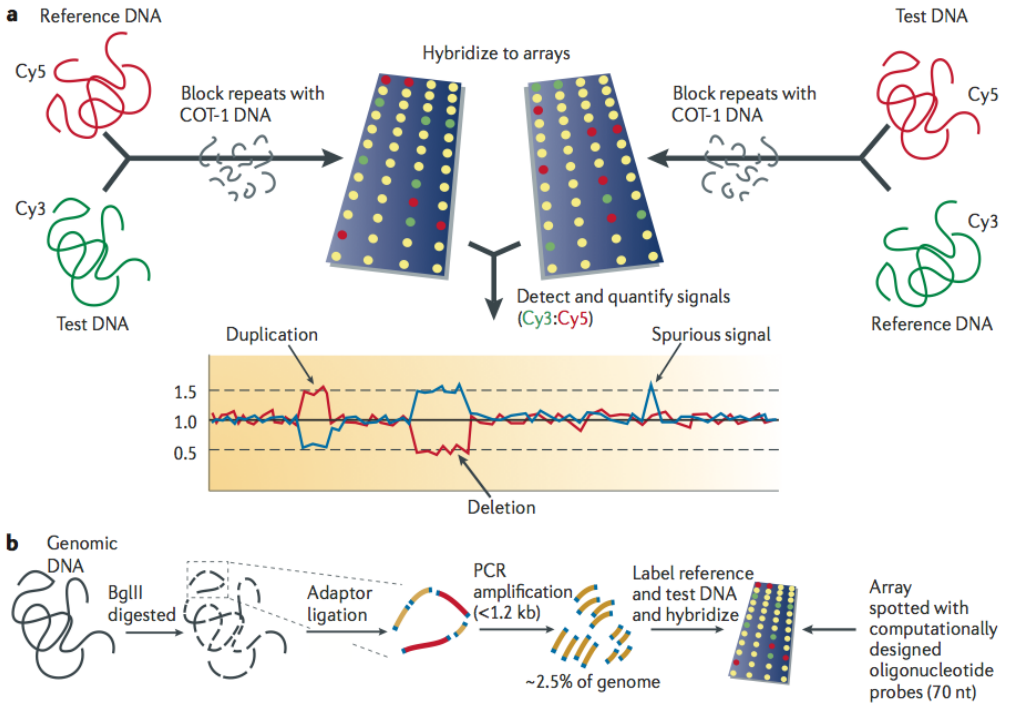


Figure 5: aCGH analysis of copy number changes. Both tumor and normal DNA are labelled with two different colored probes. The labelled DNA is than loaded on to a cDNA carrying chip. The ration of the the two intensities (in case ration of green and red) carries gain and loss information about the region on the array chip [107].

arrays only require one DNA type for the analysis.

The hybridization intensities which are obtained from spotted oligonucleotides are compared with that of controls. Any deviation in this intensity reflects a CNA change. In addition to copy number analysis SNP arrays can also give information about genotype such as loss of heterozygosity [113, 114]. For every SNP allele, 25 bases long 20 matched and mismatched probe pairs are designed. The test DNA is first restriction digested and amplified. CNA are detected by comparing intensities of matched and mismatched probes to control individuals. SNPs are placed at a median spacing of 2.5 kb in Affymetrix GeneChip 500k arrays. Because SNP arrays were used for CNA detection Affymetrix has designed SNP 6 arrays carrying around 1.8 million markers, which includes around 906,600 SNP and more than 946,000 probes for CNA

detection.

The genome wide array techniques have been very successful in determining the "CNA landscape" of cancer. A huge amount of data has been been analyzed and made available to the research community. For this thesis project all the analysis have been performed on cCGH and aCGH data which was collected and made available through Progenetix [90].

3.5.3 Role of CNA in cancer

The importance of CNA in cancer development and progression is illustrated by their presence in majority of cancer types [115, 116]. CNA do not occur in a random manner which is reflected by their re-occurring nature and a unique behavior across cancers [117, 115].

The direct effect of CNA on the expression of genes has been proven across multiple cancer types [118, 119, 120, 121, 122, 123]. In some cases CNA in a genomic region has also been associated with a global gene expression change [124]. CNA can abrogate the function of tumor suppressor genes by deleting both the alleles of the gene [125] or by deleting one copy of the gene when the other is either germline or somaticaly point mutated [35, 36, 37]. Haploinsufficiecy in some genes (e.g. PTEN [126], PTCH1, PTCH2 [127] and other DNA repair associated genes such as MSD-3 [128]) can contributes to tumorigenesis. In some cancer types the presence of CNA is an essential event in tumorigenesis [55].

CNA, particularly small amplicons, have also led to direct identification of candidate tumor suppressor genes and oncogenes [129, 130, 131]. CNA have been used to classify an individual cancer into its subtypes [132]. The presence of CNA has also been linked with treatment outcome [133].

CNA are associated with disease progression and recurrence in several cancer entities. Examples are loss of 1p, 3p,11q and gain of 17q are associated with poor clinical outcome in pediatric neuroblastoma [134, 135, 136]. Two different clinicogenetic subtypes of neuroblastoma can be defined using these imbalances, one with loss of 11q, 3p and gain of 17q and other with loss of 1q and gain of 17q. Frequently occurring gains involving 2p have been associated with advanced stages in chronic lymphocytic leukemia [133]. Gains of regions carrying genes MYC,

ERBB2, MDM2, EGFR and CCND1 are significantly associated with high grade breast cancer [137]. In breast cancer the amount of CNA is associated with tumor grade, metastasis and histology [138]. Worse clinical outcomes has been associated with an increase in overall chromosomal imbalances [139]. Analysis of CNA data has helped in defining genomic regions which are altered in multiple cancer types (e.g. gains on 1q (MCL1), loss on 3q (FHIT), gains on 8q (MYC), losses on 13q (RB1), gains on 11p (CCND1), gains on 12 (KRAS, CDK4 and MDM2) [117, 115]). CNA accumulation can start in the early phases of tumorigenesis, supporting the claim that they are required for tumor development [140, 141, 142]

CNA profiles have been used to classify cancer patients genetically. It has been shown that CNA can be used for accurate diagnosis of different malignant lymphomas and their subtypes[143]. Frequent genomic amplifications were also used to classify cancer types and it was shown that amplifications are selected according to the anatomical locations of cancer [144].

CNA are also associated with clinical phenotypes and are often used in clinics for diagnosis. Breast cancer patients with amplified ERBB2 gene are treated with Herceptin [145]. CNA gains are associated with resistance to therapies e.g. TYMS in response to 5-fluorouracil [146], DHFR in response to methotrexate [147].

3.5.4 Systems biology and CNA

CNA can spread through several genomic locations across chromosomes. Majority of the tumor cases acquire more than one CNA. These CNA can work in a cooperative manner to affect interacting gene modules. Previous work [148, 149] has illustrated some cooperative nature of CNA. [148] showed that frequently co-occurring CNA often involve interacting gene modules. However none of the work so far has accounted for the genomic complex behavior of tumor samples. Tumor complexity can be defined by the number of changes present in a tumor. Tumors samples can acquire multiple CNA. A complexity dependent behavior of these changes will help in defining co-occurring CNA landscape which is independent of the tumor complexity [150]. CNAs co-occurring in samples with overall few changes may point towards their significant

involvement in cancer.

Genes in CNA affect cellular pathways, and previous studies have tried to address the role of CNA in pathway dependencies [54, 151]. Others have integrated somatic mutation data with CNA data [152] to identify pathways altered in cancer. Some others ranked gene modules significantly affected by frequently altered genes across CNA [153]. But so far there is neither any standard method for pathway enrichment nor any large scale analysis (i.e. across multiple different entities) that has been performed on cancer CNA data. Although a single CNA can affect a multitude of genes from similar pathways the issue of genomic clustering of pathway genes [154] in relation to CNA has not been addressed yet.

Clustering has been extensively used to visualize and detect structures in cancer CNA data. Clustering of CNA patterns is being used to identify oncogenomic similarities [117, 115, 155, 156, 143] among cancer samples. Clustering of CNA data divides tumor samples into several groups with similar CNA profiles. These samples might follow a similar way of cancer progression/development. Previous work has tried to simplify the clustering by looking at frequently altered genomic regions [157, 158]. Several kinds of clustering algorithms are available to the research community, however hierarchical clustering is used more often than others as one does not need to specify the number of clusters and the results can be viewed with the help of a dendrogram tree. Multiple hierarchical clustering algorithms using different clustering algorithms are described in literature. The quality of various hierarchical clustering algorithms can be validated with the help of clustering quality measures defining the compactness, separation and connectedness among clusters. A systemic validation of hierarchical clustering can identify the best algorithm fitting the data of interest.

The application of clustering algorithms has been restricted to sample based analysis of CNA data. However until now no one has tried to address the heterogeneous and complex behavior of CNA across cancer types and use this information to find CNA separating cancer types [159].

4 Objectives and Content of the Thesis

The main objective of this thesis is "oncogenomic pattern detection in cancer copy number alteration data for pathway description and disease classification". It involves development of new tools for the analysis of cancer CNA data and than compare them to existing methods. The methods designed or available should be tested on CNA data from multiple cancer types to measure their performance. The specific aims are

1. **Development of methods for functional associations in CNA data.** This involves development of tools for various kinds of systems biology analysis on CNA data. The questions that can be addressed are

   (a) How to find co-occurring CNA in cancer samples?

   (b) Which pathways are targeted more often by CNA?

   (c) Which CNA play a role in cancer to cancer divergence, i.e. cancer classifying nature of CNA?

2. **Exploration of available methods for analysis of cancer CNA data.** While developing new methods available tools are also tested for the analogous analysis. The results obtained are compared with what has been observed before.

3. **Application of explored and developed methods to data available on Progenetix.** All the tools either developed or available should be tested on cancer CNA data available through Progenetix. The application can either involve analysis of a small data set from few cancer types or performance on a large heterogeneous data comprising of all cancer types in Progenetix.

5 Methods

5.1 Co-occurring nature of CNA - Paper I

Most of the tumors samples acquire more than one CNA during the time of cancer development and progression. Identification of co-occurring CNA can help in defining relationships among individual genomic changes accumulated in cancer. The co-altered genomic regions can harbor genes, either directly interacting with each other or playing role in co-operative pathways. Previous work has tried to address the co-occurring nature of CNA and suggested ways of identifying frequently co-altered genomic regions [148, 149]. Simultaneously altered genomic regions were shown to harbor interacting genes.

However, the co-occurring analysis of CNA is less trivial than contemplated. Cancer is a genetically heterogeneous disease and samples from one cancer type can acquire variable numbers of CNA [150]. Cancer samples can be classified on the basis of their genome complexity; "CNA complex" with a high number of CNA in contrast to samples with very few changes ("CNA simple"). The complexity of samples has to be considered while evaluating co-occurring CNA. We defined a statistical method "CDCOCA" for complexity dependence of co-occurring chromosomal aberrations, which considers the sample heterogeneity while identifying co-occurring CNA [150]. The detailed formation of the algorithm has been described in paper I.

5.2 Pathway enrichment across CNA data

Gene and pathway data

Several resources of pathway-associated annotations are publicly available (e.g. Reactome[160], Kegg[161] and Nature signaling pathways [162]). We use Nature Signaling Pathways as a source for all algorithms; it only consists of signaling cascades and does not include processes like metabolism (present in KEGG). Another reason for using Nature Signaling Pathways is its non-hierarchal pattern of annotation, contrary to e.g. Reactome where big pathways are further divided into hierarchies of pathways.

Genes were mapped to genomic locations according to Human genome version 18 (hg18/build36). Genes on chromosomes X and Y were removed prior to analysis due to inherent gender biases (e.g. prostate carcinoma) and inconsistent reporting in the CNA data sets. The remaining pathway-specific data consisted of 176 pathways containing a total of 2239 genes.

For genomic positions, we obtained the Ensembl gene list from Biomart release 54. This gene list was processed to obtain genes with unique combination of Ensembl gene identifier (Ensembl id) and chromosome start and end position. Mitochondrial and sex chromosomal genes were removed resulting in a total of 20209 genes for the input gene list.

For S-path we divided all chromosomes in to non-overlapping segments of 1Mb each dividing entire genome in to 2872 bins (we call this "segmentation"). The gene list was than mapped to the artificial segments of 1Mb and all genes for any pathway present in these bins are considered as a single gene. We expected that dividing the genome in to small bins resolved the issue of pathway genes clustering. 2239 pathway genes were mapped to 1203 unique genomic segments.

Data sets

Cancer CNA data available through Progenetix [90] is used as an input to pathway algorithms. A total of 19819 tumor samples from 132 different cancer types were analyzed for pathway enrichment. The analysis was then extended to individual cancer type for entity specific pathway enrichment profile. The genome for copy number data across all samples was reduced to

segments of 1Mb each as an input to S-path. Such an input data has been used to enrich for CNA which are altered more often than chance [163].

Model and parameters

We here define models and parameters for G-path and S-path. The only difference between G-path and S-path is the pathway input file. Genes are an input for G-path which are replaced with segments for S-path. Here we use the term genes for both; genes for G-path and segments for S-path.

A unique permutation strategy is used to compute pathway scores using G-path/S-path. Let S represents a list of samples $\{S_1, ...S_n, ...S_m\}$ of a given cancer type. A pathway P is represented as a set of genes G_i , where i is any number of genes with known genomic locations on autosomes. We consider that for pathway P more than one gene can simultaneously be affected by CNA Figure 6 shows the boxplot of number of 1Mb pathway segments altered in some pathways across 19379 tumor samples, showing on average 10% of segments are copy number altered across pathways. For every sample a pathway index $SP_n = N_P$ is obtained. N_P represents number of genes for pathway P which are copy number altered in sample S_n, $N_P \, \epsilon \, G_i$. The overall pathway score for pathway P across all samples is obtained as $OP_P = \sum_{n=1}^{m} SP_n$. The list of overall pathway scores is represented as $\{OP_1, ...OPp, ...OPs\}$for s pathways. The corresponding list obtained on permutations is represented as $\{OP_1^*, ...OPp^*, ...OPs^*\}$. OP_p can vary between 0 (when no gene for pathway P is copy number altered) and $\max\{i * m\}$(when all genes i are copy number altered across all m samples). All the three algorithms (G-path, S-path and H-path) identifies pathways disapproving the null hypothesis that "there is no association between pathway genes and CNA".

Pathways enrichment with G-path and S-path

In this section we describe the two algorithm G-path and S-path. Genes for G-path represent same as segments for S-path.

1. Overall pathway scores for all pathways $\{OP_1, ...OP_p, ...OP_s\}$are computed.

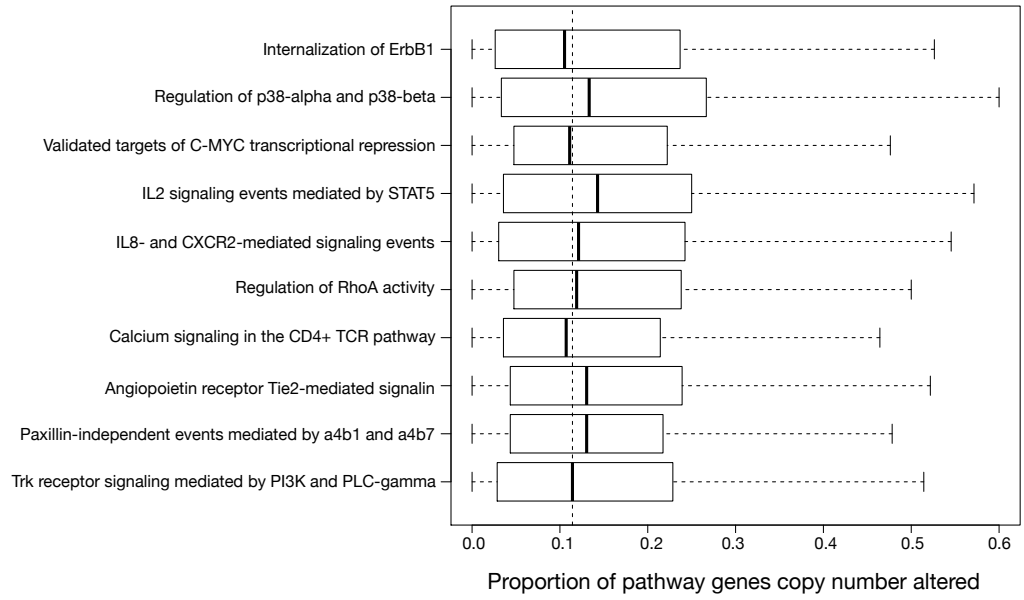


Figure 6: Proportion of genes showing CNA across entire cancer CNA data; 10 randomly chosen pathways are shown here. The vertical line represents the average number of genes being hit (median of per-case fraction of window bins overlapping CNA). The overall fraction of window bins being hit in a given pathway uctuates around the expected value.

2. A null vector $C = \text{NULL}$ of length s is defined. C measures the index how many times expected scores across pathways are greater than or equal to observed scores.

3. For permutation all Ensemble genes (2872 segments for S-path) are randomly distributed on to new genomic locations. This permutation helps in keeping gene pathway membership consistent while affecting only pathways Vs genomic location membership.

4. Expected pathway score $\{OP_1^*, ...OP_p^*, ...OP_s^*\}$on permutation for all pathways is computed.

5. If expected score for pathway P is greater than or equal to the observed score C_p is incremented as

$$C_p = \left\{ \begin{array}{ll} C_p + 1 & if\ OP_p^* >= OP_p \\ C_r & if\ OP_p^* < OP_p \end{array} \right\}$$

6. Step 2 to 5 are repeated over Z number of times (Z = 10000 in current analysis) increasing the counter C for pathways which are altered by chance.

7. After all permutations a p value vector $\{p_1, ..p_p, ..p_s\}\ \epsilon\ P$ are caclulated from C as $p_p = \frac{C_p}{Z}$

8. All the p values are corrected for false discovery rate using Benjamini Hochberg correction.

H-path Algorithm

In this section we describe the methodology used by [152]. The null hypothesis tested for is that pathway gene membership is random. For pathway P with G_i genes where i is any number of genes with known genomic locations on autosomes. An index P_h is computed as number of samples having more of more gene altered for pathway P. For simulation G_i are sample from all set of pathway genes without replacement and index P_h^* is recomputed. The p value for the pathway P with index P_h is computed as number of times $P_h^* >= P_h$. All p values are then FDR corrected using Benjamini-Hochberg correction.

Genomic clustering of pathways

Genes for cellular pathways are clustered on the genome [154]. For all the pathways a clustering score is compute as described in [154]. Since for S-path genes from pathways are represented as segments the pathway clustering score is computed using these segments and not genes. For each gene pair on same chromosome from a pathway a pair wise score of clustering is computed as

$$pair\ score = \frac{average\ length\ of\ chromosomes\ in\ genome}{distance\ between\ genes}$$

For genes on different chromosome distance is computed as

$$pair\ score = \frac{average\ length\ of\ chromosomes\ in\ genome}{average\ length\ of\ chromosomes\ the\ genes\ are\ located\ on}$$

The pathway clustering score is the sum of all pair wise scores divided by the number of genes in that pathway. The original clustering score is compared to the expected scores obtained by randomly creating pathways with same number of genes and then recalculating the expected pathway clustering score (10000 permutations). p-values are generated by comparing the original and expected scores as

$$pvalue = \frac{number\ of\ times\ expected\ clustering\ score \geq observer\ clustering\ score}{number\ of\ permutations}.$$

The p values are corrected for false discovery rate using Benjamini Hochberg correction.

5.3 Evaluation of hierarchical clustering on CNA data

Hierarchical clustering is very commonly used to find heterogeneous groups in data. The aim was to find best hierarchical clustering methods to cluster CNA frequency profiles across 160 different cancer types.

Data

Cancer CNA data consisting of chromosomal and array CGH experiments across 160 cancer types, and a total of 25579 cancer samples was used. All cancer samples were classified on the basis of International Classification of Disease Code (ICD). For validation of clustering methods CNA information was reduced to 80 genomic intervals covering the entire genome. Gain and loss events on every locus were considered separately.

Clustering validation

Internal cluster validation methods provide a way to determine the quality of clustering. Three different cluster validation methods were used; They are silhouette width, Dunn index and connectivity to find the best hierarchical clustering method for this data.

Silhouette width computes the compactness and separation among clusters. Well clustered observations have a value close to 1 and poorly clustered close to -1. For any observation i

$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$

where a_i is the average distance between i and all possible observations in same cluster and b_i the average distance between i and observations in the nearest neighbor cluster

$b_i = min_{C_k \in K, C_i} \sum_{j \in C_k} \frac{dist(i,j)}{n(c_k)}$

where C(i) is the cluster with observation i, dist(i,j) is the distance between observations i and j and n(C) is the cardinality of cluster C. K is is clustering partition of k clusters. Silhouette width lies between -1 and 1.

The connectivity measures the compactness, separation and connectedness of a clustering algorithm. It can be obtained as

Connectivity = $\sum_{i=1}^{N}, \sum_{j=1}^{L} x_{i,nn_{i(j)}}$

Where N is the number of observations, $nn_{i(j)}$ is the jth nearest neighbor of i and $x_{i,nn_{(j)}}$ is zero if i and j are part of same cluster otherwise 1/j. Connectivity can assume a value between 0 and infinity and should be minimized.

Dunn Index is the ration of the smallest distance between observations not in same cluster to the largest intra-cluster distance.

$D(K) = \frac{min_{C_k, C_l \in K, C_k \neq C_l}(min_{i \in C_k, j \in C_l} dist(i,j))}{max_{C_m \in K} diam(C_m)}$ where $diam(C_m)$ is the maximum distance between observations in cluster C_m. Dunn index varies between zero and infinity, and it should be maximized.

The known cluster validation methods were compared to a new method "Tree length statistics (TLS)" to find the best validation method among them. The TLS uses the information present in the dendrogram tree obtained on clustering to determine the quality of clustered data. TLS is defined as the sum of all parent-child distances in the dendrogram tree (Figure 7). Tree distance between two cancer types on a dendrogram tree, obtained using clustering of cancer CNA frequency profiles reflects the discrepancy in between CNA profiles of these cancers. For any node i, the tree height difference between this node and its immediate parent j is measured as $TH_j - TH_i$. The overall tree height of a tree with n nodes is than obtained as $OTH = \sum_{i=1,j=1}^{i=n,j=n} TH_j - TH_i$ (Figure 7).

A permutation strategy was used to identify the hierarchical clustering algorithm in combination with a cluster validation method fitting best to the data. The strategy was as follows

1. Cluster CNA frequency profiles across all genome intervals across all cancer types.
2. A counter C is set to zero.
3. Calculate the observed cluster validation index (OVI).
4. Randomize the frequency values among all cancer types to generate a new frequency matrix with similar number of cancer types and genomic intervals.
5. Cluster the randomized CNA frequency profiles and compute the random cluster validation index (RVI).

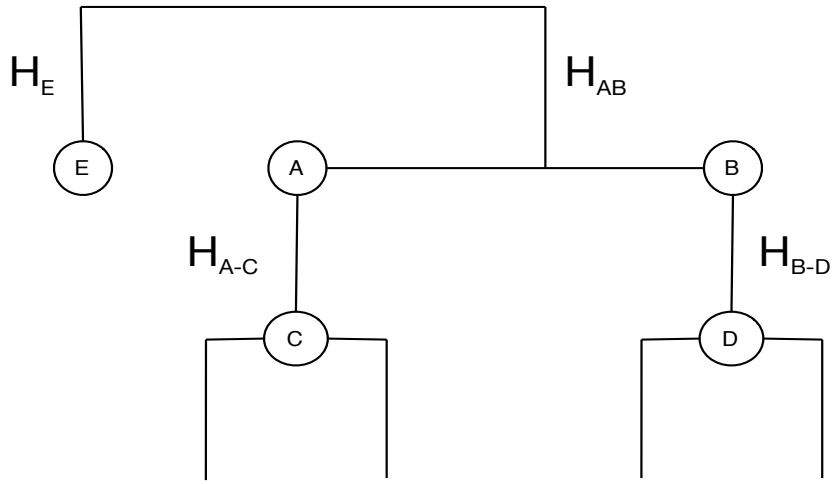


Figure 7: Schematic representation of the summed branch-length tree height statistic. Overall tree height is computed by summing up the distance between all parents and child nodes. Note that the branch lengths of terminal branches ("leafs") are not considered. Overall tree height $= H_{A-C} + H_{B_D} + H_{AB} + H_E$.

6. If $RVI >= CVI$, C is incremented as, $C = C + 1$.

7. p-value for the given cluster statistic, at the end of N (1000) permutations is computed as $p = \frac{C}{N}$.

5.4 Non-neutral CNA - Paper II

Clustering of CNA data has mostly been biased towards sample based analysis involving frequently altered genomic regions. We aimed at identification of genomic regions playing a crucial role in clustering of cancer types. It is observed that the CNA landscape is very heterogeneous and with greatly varying average CNA frequencies per cancer type (Paper 2). When CNA frequency profiles among cancer types are used for analysis using clustering, results may be distorted depending on the entities analyzed. We focused on identification of genomic regions playing a crucial role in clustering of caner frequency profiles. For this normalized CNA frequency profiles of cancer types were clustered using hierarchical ward clustering. A simple permutation based methodology was used to measure the contribution of individual genomic region to clustering of cancer types (Paper 2). A new clustering quality index termed "Tree height statistics" was used as a determine the quality of clustering obtained by permeation of genomic regions. All the methodology is defined in detail in Paper II.

6 Results and Discussion

The fact that CNA are important for cancer has been well established by their presence in all tumor types. Because of their importance, an advancement in technologies for their easy identification has been made. Several data sources such as SKY-CGH [100], Progenetix [90] and arrayMap [164] have collected and made this data available to the research community. Because of large scale data availability there is in immediate need for detailed and systemic analysis of CNA data.

6.1 Co-occurring nature of CNA - Paper 1

The overall genome involved in CNA across cancer types vary considerably (Figure 8) which is affected by the number of CNA present in samples. Samples can be classified on the basis of their CNA complexity which can be defined through the number of CNA present in a sample and/or the overall amount of genome involvement. Cancer samples can be "CNA complex" (a high number of CNA) or "CNA simple" (very few changes). This complexity of samples has to be considered while evaluating co-occurring CNA. A statistical method "CDCOCA" (complexity dependence of co-occurring chromosomal aberrations) considering the sample heterogeneity while identification of co-occurring CNA was formulated([150]). CDCOCA was applied to two cancer types; mantle cell lymphoma and bladder carcinoma. It was observed that most of the co-occurring changes across these two cancers resulted from a background of multiple and extended CNA. While correcting for sample complexity most of the associations involving high frequent changes were removed indicating that these high frequent changes may be related to an overall high genomic instability.

One can either focus on frequently co-occurring CNA, which can be identified using modified version of CDCOCA- "CICOCA" (complexity independence of co-occurring chromosomal aberrations) or CNA from cancers samples with less complex behavior (CDCOCA). CICOCA and previous work focussed on identification of frequently co-occurring CNA however, with the help of CDCOCA one can test for the specificity of their association. Most of the associations

obtained using CICOCA and previous work [149] were also observed to be enriched using CDCOCA. An advantage of CDCOCA is that it requires only one p value cutoff, unlike other methods [148, 149] also requiring frequency based thresholding. This helps in identification of both frequently and eminently associated events of CNA. Genes from cancer associated pathways could be located in the enriched co-occurring regions. This bestowed the motivation for further analysis, if genes from analogous pathways are targeted by CNA i.e. some pathways genes are preferentially enriched in cancer CNA.

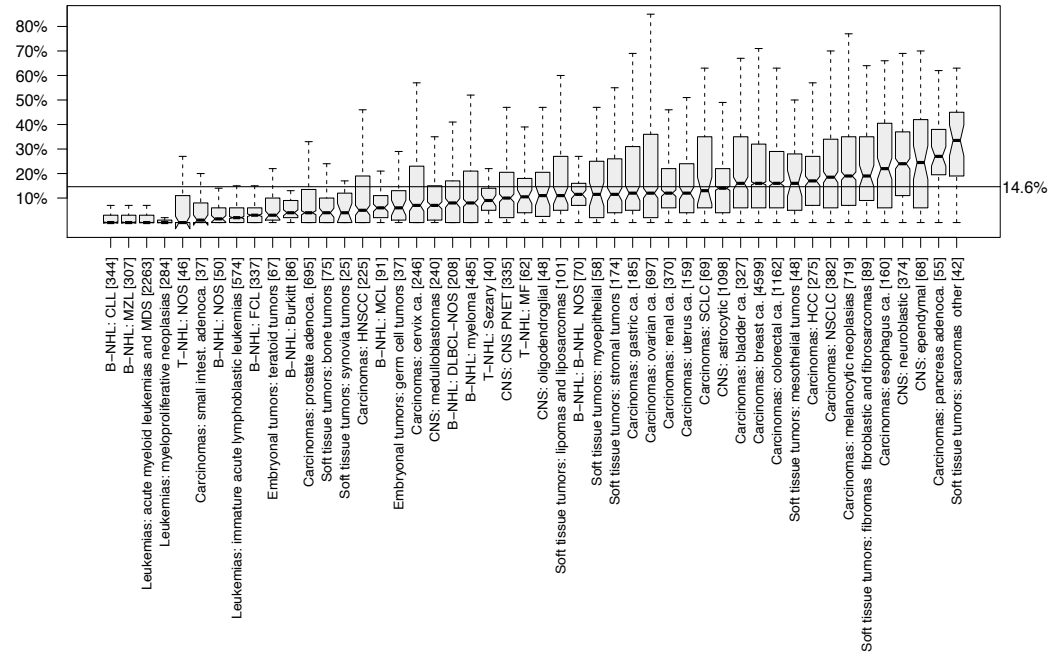


Figure 8: Box plot for the percent of genome involved in CNA across cancer entities represented in Progenetix [90].

6.2 Pathway enrichment across CNA data

CNA can target multitude of genome genes and cellular pathways. A pathway centric analysis involving CNA data can help in defining cancer associated pathway modules targeted more often than chance. This analysis can be more useful than a gene centric analysis firstly because most of the CNA cover large genomic regions and targeting high number of genes at once and secondly for a similar pathway alteration multiple genes can be selectively/differentially targeted in a cancer. Pathway analysis using CNA data is non-trivial as several issues such as genomic clustering of pathway genes and multiple pathway genes are targeted by a single CNA. All these issues have to be considered for a CNA specific pathway analysis.

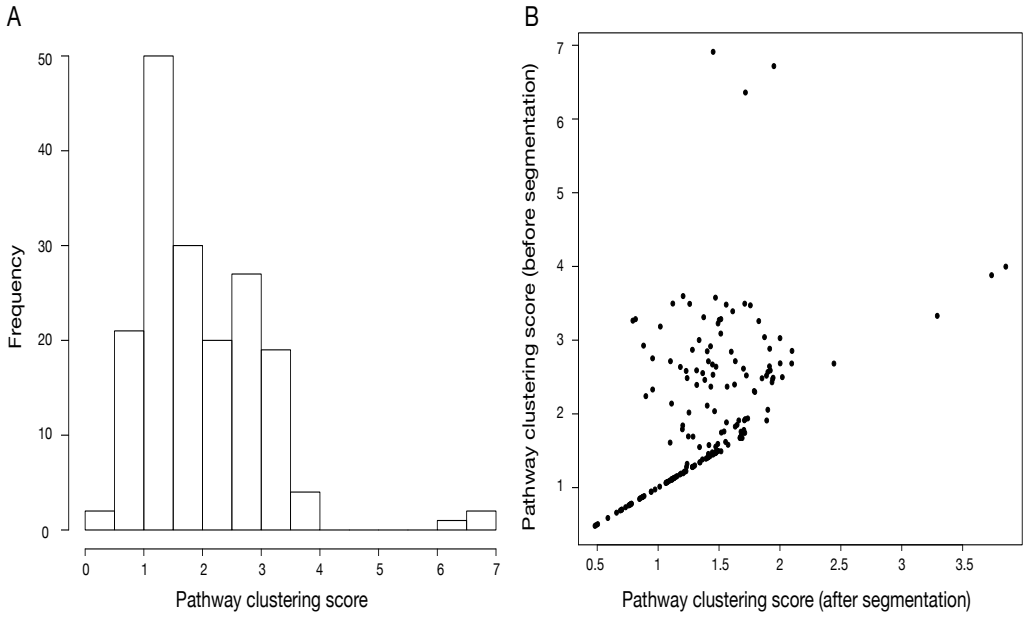


Figure 9: A) The distribution of scores obtained for the genomic clustering of pathway genes. log10 of the scores are plotted here for simplicity. B) Comparison of pathway clustering score before and after segmentation. Dividing the genome into small segments reduces the pathway clustering scores.

Clustering of pathway genes

The pathway clustering score followed nearly a normal distribution with data having a mean score of 89424.79 (Figure 9). The clustering score was independent of the number of genes in a pathway (Figure 10). Dividing the genome into artificial segments of 1Mb resulted in a reduction of genomic clustering scores (Figure 9) and the mean clustering score was reduced to 114.465. Also genomic pathway clustering signal was lost after segmentation (Figure 10). Before segmentation genes from 47 pathways were clustered on genome with an FDR of 0.15. However with segmentation none of the pathways remained clustered (Figure 10). This analysis points towards the key issues about considering genomic segments as a set of pathway components and not genes.

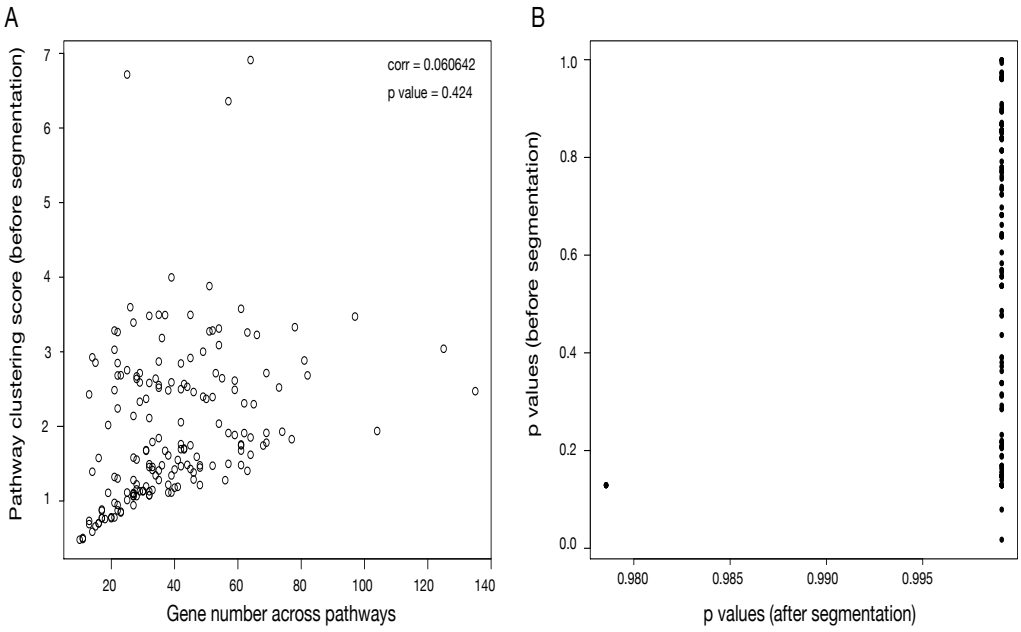


Figure 10: A) Plot showing that clustering of pathways is not affected by the number of genes in that pathway. B) Benjamini Hochberg corrected p values for pathway clustering score before and after segmentation are plotted here. None of the pathways were clustering on genome after segmentation.

Comparison H-path, G-path and S-path

All the three methods were used to obtain a list of pathways targeted more often than chance by CNA. Cancer CNA alterations span through multiple pathway genes (Figure 6), indicating the importance of the total score as a pathway score parameter. A maximum FDR cutoff of 0.15 was used as a selection criteria to find enriched pathways. H-path led to an enrichment of only one pathway (Figure 11) whereas with G-path 4 pathways (one genomic unclustered and three clustered) were enriched (Figure 11). Using S-path 17 pathways were found to be significantly altered (Figure 11). Genes from five of these pathways were clustered in genome however since segments were considered this genomic clustering was not significant and did not affect the results.

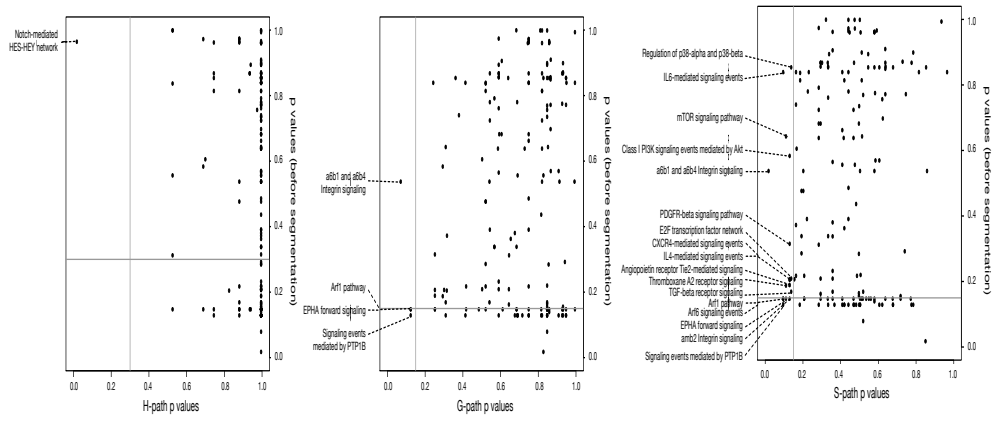


Figure 11: Plot of p values of pathways enriched using A) H-path, B) G-path, C) S-path. Grey lines indicated the maximum FDR of 0.15

All the pathways enriched using G-path were also enriched using S-path along with additional pathways identified with similar FDR cut-offs. The issue of genomic clustering is also resolved by S-path as all the genes present in a window bin of 1Mb are considered as a single event. H-path efficiently identifies pathways which are hit more often with CNA than by chance. However the analysis lacks the ideology that for a given pathway more than one gene can be

simultaneously copy number altered. The permutation strategy used by G-path and S-path is unique in itself as till now no one has considered to keep genes (segments for S-path) and pathway membership constant on permutations. Another advantage of G-path/S-path is that a larger range of background CNA are considered in analysis.

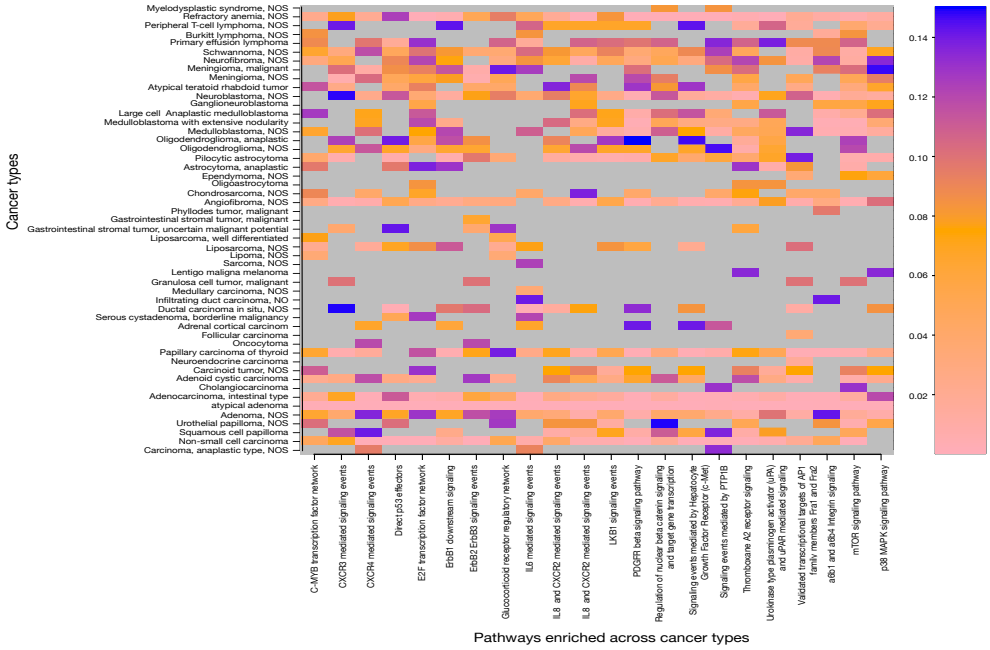


Figure 12: Pathways enriched in more than 20 cancer types with S-path are represented here. 51 (51/59). cancer types showed an enriched of commonly altered pathways. The BH corrected p values are color coded as shown in right.

Pathways enriched

We extended the analysis using S-path to CNA data from 132 different cancer types. 59 cancer types showed an enrichment of at least one pathway with an FDR cut off of 0.15. 22 pathways were enriched in more than 20 individual cancer types (Figure 12). On further scrutiny we nailed down to 8 signaling pathways (Figure 13), which were identified in integrative analysis as well

as were present in more than 20 ICD types. Interesting candidates included mTOR signaling pathway and Thromboxane A2 receptor signaling, both of these pathways were enriched in more than 30 cancer types. Identification of these pathways predicts an increased proliferation or antiapoptotic properties in cancer cell (Figure 13).

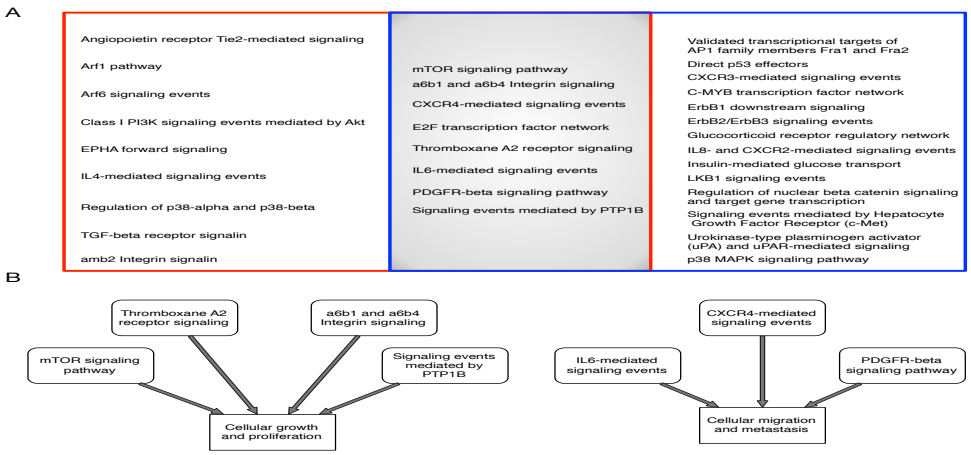


Figure 13: Pathways enriched by a combinatorial analysis are shown in the red rectangle and the ones enriched in multiple cancer types (20) are shown in the blue rectangle. In the intersection of both rectangles (grey shaded region) are pathways which were looked at in details. B) Network of pathways enriched in cancers. Different enriched pathways (shown in rectangle with round corners) converge to affect cancer-associated cellular responses (rectangle with smooth corners).

6.3 Best hierarchical clustering

Hierarchical clustering is very commonly used to find heterogeneous groups in data. The aim of current analysis was to find best hierarchical clustering algorithm fitting the data of interest with the help of cluster validation methods. Simultaneously it was also tested which cluster validation method gives the best cancer to cancer separation on a dendrogram tree.

Four cluster validation methods (3 known and one new; TLS) were used to find the best hierarchical clustering algorithm fitting CNA data. All the internal cluster validation measures require the data to be divided into different clusters. This was obtained by cutting the dendrogram tree at various heights to generate desired number of clusters.

Connectivity measures connectedness among clusters i.e. to which extent observations are in the same clusters as their nearest neighbors and its value should be minimized. The best clustering on the original data in comparison to random data sets was observed using hierarchical ward clustering with 8 separate clusters (Figure 14d, p value 0.008). With hierarchical ward clustering connectivity increased with an increase in number of individual clusters (Figure 14).

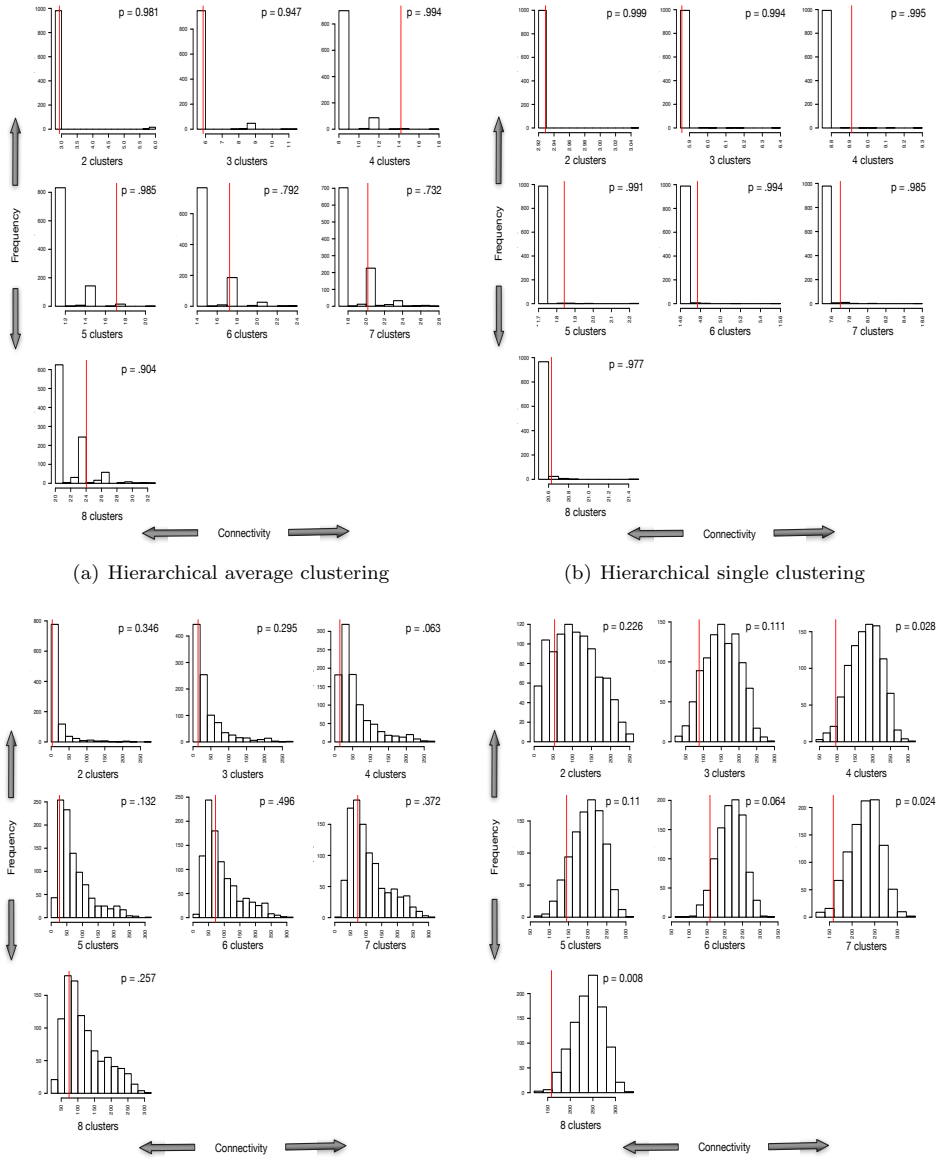
In comparison to connectivity, Dunn index and silhouette width measure the compactness of clusters. Both these measures look for intra-cluster variance to assess cluster homogeneity. The value for Dunn index should be maximized whereas silhouette width should be close to 1 for perfect clustering.

Dunn index classified hierarchical single, average and complete clustering with 2 separated clusters as the best clustering algorithm (Figure 15, p value = 0.003). In general Dunn index followed an inverse relation to the number of clusters. When focussed on silhouette width all the clustering algorithms worked equally well (Figure 16) on the original data in comparison to random datasets. However the best clustering was observed using hierarchical single and average clustering with a maximum silhouette width of 0.6 for 2 separated clusters.

When focused on the new cluster quality method TLS, all clustering algorithms worked better on original data than on random ones(Figure 17). However hierarchical ward clustering on original data gave the a maximum separation of 87.13 SD (standard deviation) than the mean

of random data. When compared to other internal cluster validation methods TLS outperformed all of them hence it should be used as a cluster quality index. TLS doest not depend on the number of final clusters so it does not require the tree to be cut at random heights to produce desired number of clusters unlike in other internal cluster validation methods.

With this analysis it was observed that hierarchical ward clustering generates best separation of cancer types on a dendrogram tree. TLS was the best clustering validation method. The combination of ward clustering and TLS was used to find non-neutral CNA in accumulated cancer CNA data.



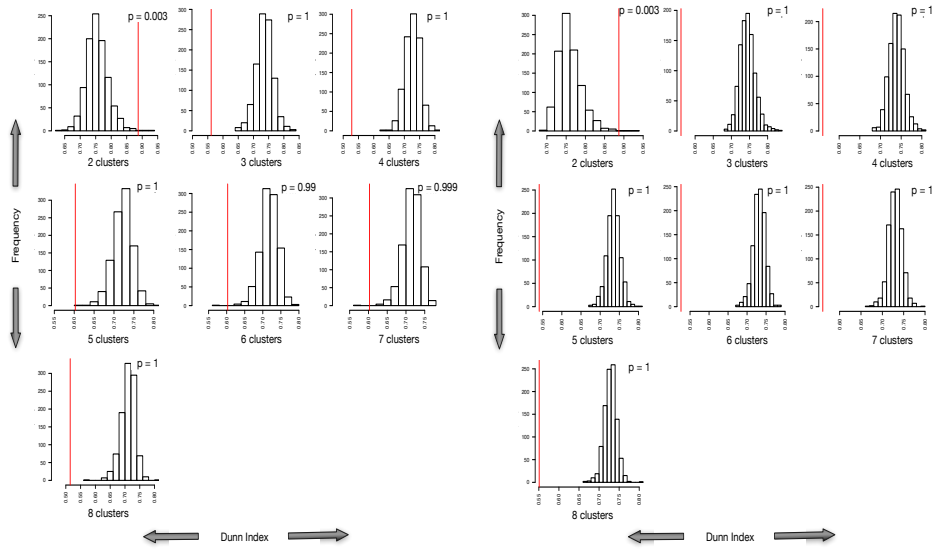
(a) Hierarchical average clustering

(b) Hierarchical single clustering

(c) Hierarchical complete clustering

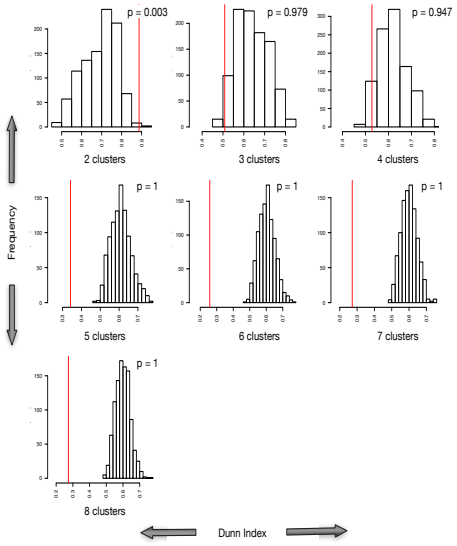
(d) Hierarchical ward clustering

Figure 14: Measure of connectivity using different hierarchical clustering algorithms. The connectivity values obtained by randomizing the data are represented using histograms bars. The observed connectivity on the original data is shown with the help of red line. The p values on right represent the how good the original clustering was than on a randomized data.

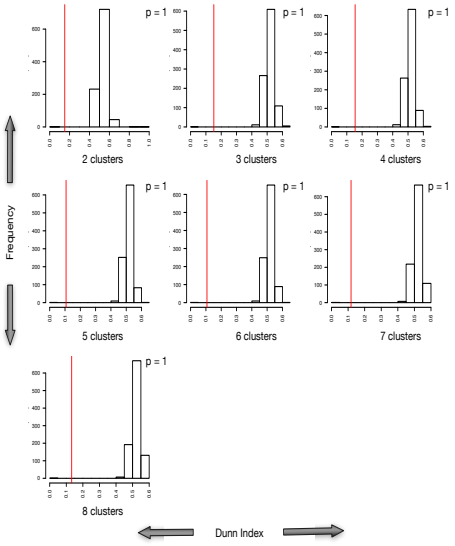


(a) Hierarchical average clustering

(b) Hierarchical single clustering

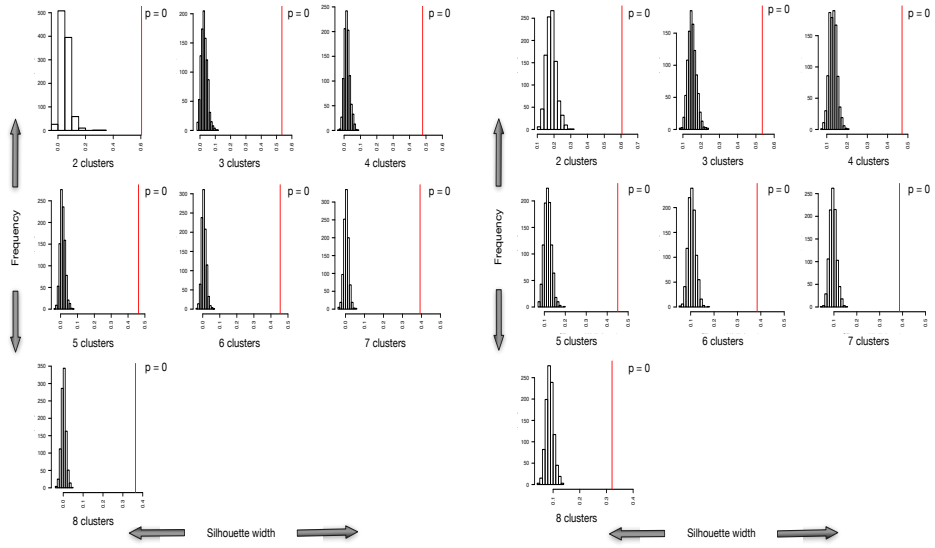


(c) Hierarchical complete clustering



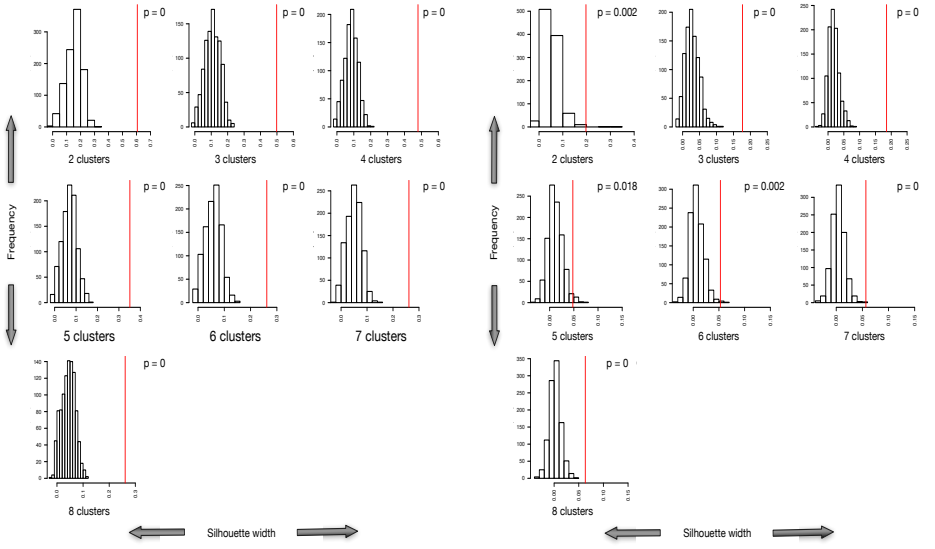
(d) Hierarchical ward clustering

Figure 15: Measure of Dunn index using different hierarchical clustering algorithms. The Dunn index values obtained by randomizing the data are represented using histograms bars. The observed Dunn index on the original data is shown with the help of red line. The p values on right represent the how good the original clustering was than on a randomized data.



(a) Hierarchical average clustering

(b) Hierarchical single clustering



(c) Hierarchical complete clustering

(d) Hierarchical ward clustering

Figure 16: Measure of silhouette width using different hierarchical clustering algorithms. The silhouette width values obtained by randomizing the data are represented using histograms bars. The observed silhouette width on the original data is shown with the help of red line. The p values on right represent the how good the original clustering was than on a randomized data.

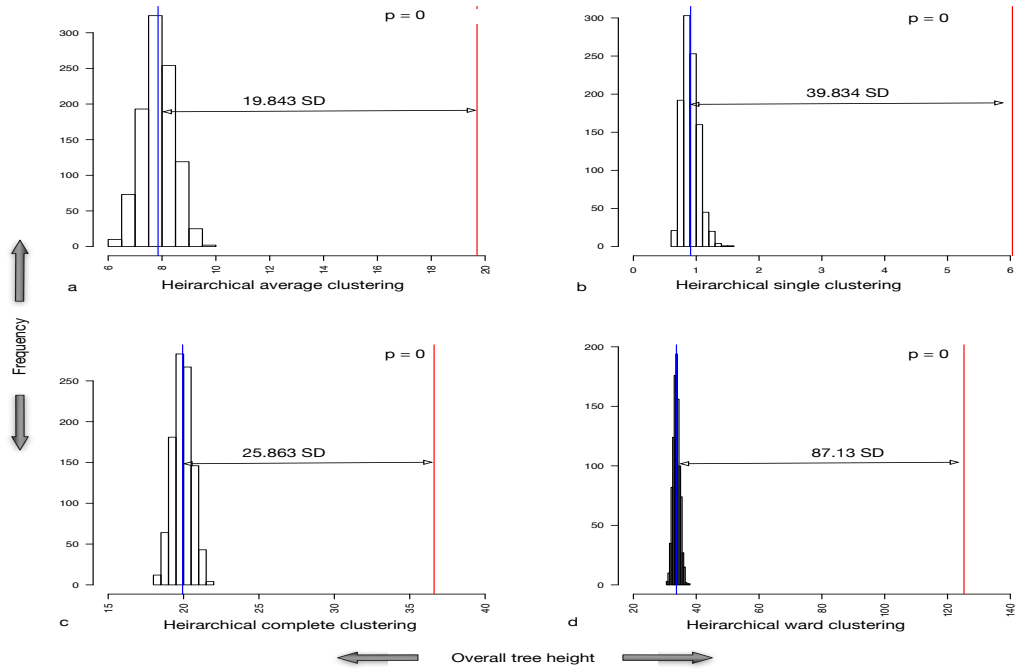


Figure 17: The distribution of overall tree height obtained by randomizing the data is plotted using histogram bars. The overall tree height on original data is represented with red line. The mean overall tree height obtained using randomized data is shown with blue line. Ward clustering (d) gave the maximum separation of original tree height than mean of randomized data (88.13 SD)

6.4 Non-neutral CNA - Paper II

CNA can also be used to identify oncogenomic similarities among cancer samples. Clustering methods are used as a way to visualize such similarities and find groups in data [117, 144, 157, 156, 143]. However most of the clustering has been performed only on samples, with few exceptions [117, **?**]. With the availability of huge amount of CNA data we aimed at clustering cancer types (not samples) and to identify genomic regions playing a crucial role in cancer type specific divergence. For this a dataset of 25579 samples from 160 different cancer types, classified on the basis of ICD-O-3 code was used.

It was observed that overall sample to sample heterogeneity results in a greatly varying average frequencies of CNAs per cancer type. The variation in overall CNA occurrence frequencies among genomic intervals should be considered before clustering analyses. In our analysis the aggregated frequency profiles across cancer types were normalized to the overall mean observed across the entire data set. The clustering of cancer types using normalized frequencies produced a signal resulted in grouping of cancer types with similar "root" cell types together. This clustering was better than expected by chance.

With the confirmation of a signal in the dendrogram tree we aimed on finding genomic changes playing a role in cancer to cancer divergence, termed as "non-neutral". We used the dendrogram tree obtained using clustering of normalized frequencies to look for non-neutral CNA. When several internal clustering measures were compared to a new method "TLS", using a randomization approach, it was observed the best separation among cancer types was observed using hierarchical Ward clustering. TLS outperformed other internal cluster validation methods. TLS was than used to identify changes playing a pilot role in cancer to cancer divergence. The tree length statistic was ideal for measuring the quality of clustering as it does not require to cut the tree at several heights to produce clusters of variable size, which is a prerequisite for several external and internal cluster validation methods. 43 genomic regions (out of 160) were identified to be enriched in the analysis (corrected p value ≤ 0.016). There was no preferential bias towards losses or gains, and very few (14 out of 43) changes overlapped with the frequently

altered regions pointing that analysis using only frequently altered regions may not always be very useful in aiding data-driven evaluation of cancer (sub-) types.Changes such as 7q losses (present in myeloid and germ cell tumors) and 8q losses (present in meduloblastomas) were identified as targets not reported before to be selectively present in these cancer types. The analysis was done only using CNA data reduced to band resolutions (80 genomic intervals) so no claims could be made about the genes present in this region. Long regions have been reported to be frequently altered in cancer types [115].

A detailed description of methodology and results can be obtained in Publication III, section 10.3.

7 Conclusion

Cancer is the second dominating cause of death in the world. It is a group of diseases associated with genomic changes. Genomic changes can occur during the normal phase of cell cycle or can be induced by external factors such as radiation damage and inflammation. CNA are one type of genome change observed in majority of neoplasias. Analysis involving CNA data can prove to be very useful to determine the role CNA play in cancer development and progression. This Ph.D. thesis is focusses on the systematic analysis of CNA data.

A new statistical method to find co-occurring CNA in cancer was formulated. CDCOCA efficiently identifies changes with an overall less complex CNA background by considering the number of CNA across tumor samples. When CDCOCA was applied to two different cancer datasets it was observed that most of the frequent CNA were removed from the results reported, indicating that the majority of highly frequent changes were related to a high genome instability. While looking at co-occurring nature of CNA one can focus either on frequently co-occurring or co-occurring CNA from a less complex background. When complexity is considered, CNA associations may point towards preferred pathways modified during the events of carcinogenesis. CDCOCA can prove to be very useful for defining associations at gene level. Thus, the results obtained with the method can be further explored using known protein-protein interactions.

Systematic analysis of genomic CNA can provide information about genes involved in cancer initiation and progression. Recurring CNA have been used in identification of candidate oncogenes. However, recently the complex contributions of multiple, cooperating genes and pathways has come into focus. With the commonly observed involvement of large genomic regions, CNA can affect the expression of a multitude of genes far beyond canonical oncogenes or tumor suppressors. CNA can be used to identify the functional cooperative gene modules targeted by them. We used a new approach to identify pathways whose genomic loci are enriched among CNA. Our analysis was purely focussed on CNA data which is available for a large number of tumor samples.

A detail analysis of some tools which can be used to find CNA targeted signaling pathways

in cancer is presented here. According to the analysis, CNA harbor several genes and can in turn affect several pathways. When the genomic clustering of pathway genes is considered then a gene based analysis to look for enriched pathways does not produce any significant results (H-path). The segmentation approach (S-path) identifies pathways altered in CNA data and considers the fact that these pathways can be clustered on the genome. However, it does not solve the issue of identifying which genes could be important for a clustered set of genes for any pathway. In our analysis of various unrelated cancer types, we were able to identify common signaling pathways that likely have a pathogenetic relevance in these entities. The results of our analyses identified a convergence of distant signaling pathways towards a restricted set of interconnected cellular responses.

To test which clustering algorithm along with a cluster quality measure gives the best separation of cancer CNA frequency profiles than of a random one, it was observed that TLS with hierarchical ward clustering was the best fit to our data. When cancer types were clustered using their normalized CNA frequency profiles, we identified a new category of CNA playing crucial role in cancer to cancer divergence. This analysis disapproved the point that only frequent CNA can be used in clustering cancer data. With our current study, involving 160 different cancer types, we aimed to provide a generalized approach for identification of changes relevant for genesis of individual cancer types. Our methodology might prove to be useful in separating biological related entities.

Most of our analysis is based on low resolution data from chromosomal CGH experiments, collected through publications. With the recent availability of high resolution (www.arrayMap.org [164]) and sequencing data our tools can prove to be very useful in defining cancer to cancer genome signature. Introduction of other kinds of genomic data such as somatic mutations, methylations, and expression can help in validating (simultaneously extending) the results procured from CNA data to other kinds of alteration reported in cancer. This might further indicate if CNA affect cancer in similar way as other kinds of genomic changes and target same gene modules.

8 Outlook

Analysis of cancer genomes can prove to be crucial for finding genes and pathways responsible for cancer development. Currently the largest amount of data available through cancer genome mutation studies is CNA data. I have developed tools for the analysis and interpretation of this data. With advancement in molecular screening technologies different kinds of whole genome mutation data (such as somatic mutations, methylation) will be made available to the research community. While consortiums such as TCGA (The cancer genome atlas) and ICGC (International cancer genome consortium) are playing a pilot role in such an analysis, an important aspect of these projects is the generation of rich datasets which can be included into meta-analysis studies.

We believe that input from different kinds of genomic mutation data can increase the prediction power of our algorithms. None of the research so far has been focussed on a combinatorial analysis involving all kinds of mutation data across multiple cancer types. A combined analysis of cancer genome data can be used to answer questions such as

1. **Preferential bias of genes towards any kind of genomic alterations** - Whole genome somatic mutation data can be used to find, if there is a preferential bias of genes towards different genomic mutations. This analysis can help in concluding if some genes are preferred to be targeted by any specific genomic change or are in general targeted more often by any type of change.

2. **Co-occurring and mutually exclusive gene modules** - Our in house developed algorithm CDCOCA has proven to be useful to find co-altered genomic CNA regions. Our algorithm can be used to find out genes which are either altered together (co-occurring) or are mutually exclusive across entire mutational landscape of cancer. This analysis will help in identifying key gene associations in cancer. Genes altered together more often than chance can be essential for cancer development or progression whereas genes showing a mutually exclusive behavior would define different genomic pathway by which a cancer can develop.

3. **Integrative analysis using other data types** - All of our analysis done till now is being focussed only on cancer CNA data. Integration of RNA and protein expression can help in identification of CNA regions altering either (or both) RNA expression or protein expression. This can help in reducing the complexity of genes altered by CNA to identify the probable candidates which show a standard pattern from CNA $\rightarrow$ RNA expression $\rightarrow$ protein expression.

9 Publications and Manuscripts

9.1 Publication 1: CDCOCA: A statistical method to define complexity dependence of co-occuring chromosomal aberrations (Published)

Kumar et al. BMC Medical Genomics 2011, 4:21
http://www.biomedcentral.com/1755-8794/4/21

BMC Medical Genomics

RESEARCH ARTICLE Open Access

CDCOCA: A statistical method to define complexity dependence of co-occuring chromosomal aberrations

Nitin Kumar[1], Hubert Rehrauer[2], Haoyang Cai[1], Michael Baudis[1*]

Abstract

Background: Copy number alterations (CNA) play a key role in cancer development and progression. Since more than one CNA can be detected in most tumors, frequently co-occurring genetic CNA may point to cooperating cancer related genes. Existing methods for co-occurrence evaluation so far have not considered the overall heterogeneity of CNA per tumor, resulting in a preferential detection of frequent changes with limited specificity for each association due to the high genetic instability of many samples.

Method: We hypothesize that in cancer some linkage-independent CNA may display a non-random co-occurrence, and that these CNA could be of pathogenetic relevance for the respective cancer. We also hypothesize that the statistical relevance of co-occurring CNA may depend on the sample specific CNA complexity. We verify our hypotheses with a simulation based algorithm CDCOCA (complexity dependence of co-occurring chromosomal aberrations).

Results: Application of CDCOCA to example data sets identified co-occurring CNA from low complex background which otherwise went unnoticed. Identification of cancer associated genes in these co-occurring changes can provide insights of cooperative genes involved in oncogenesis.

Conclusions: We have developed a method to detect associations of regional copy number abnormalities in cancer data. Along with finding statistically relevant CNA co-occurrences, our algorithm points towards a generally low specificity for co-occurrence of regional imbalances in CNA rich samples, which may have negative impact on pathway modeling approaches relying on frequent CNA events.

Background

Genetic alterations are an absolute requirement for malignant neoplasias in humans [1,2]. Both kind of genetic alterations and order of occurrence are important for cancer development and progression [3]. Additionally to sequential event models, large scale analysis of genomes from patient's tumors have shown that multiple genetic abnormalities can promote the development of one cancer entity [4]. Alterations in cancer genome can range from subtle sequence changes (e.g. point mutations) over structural alterations with functional impact on the coding sequence (e.g. generation of fusion genes by chromosomal translocations) to regional or whole-chromosome copy number abnormalities (see e.g. [5-7]).

Through a gene dosage effect, genomic copy number alterations (CNA) may lead to insufficient expression of tumor suppressors or overexpression of proto-oncogenes, respectively. Recurrent CNA have been identified in nearly all cancer entities [8-10]). Comparative Genomic Hybridization (CGH) [11,12] is a genome wide CNA screening technology which has been widely applied throughout the last two decades. Building on the reverse in situ hybridization principle developed for chromosomal CGH [13], genomic microarray technology (aCGH; [14,15]) now utilizes intensity values from up to millions of short DNA sequences to derive regional copy number estimates.

Large data sets from copy number screening experiments should provide a powerful resource for oncogenomic data

* Correspondence: mbaudis@gmail.com
[1]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich, Switzerland
Full list of author information is available at the end of the article

mining studies. In contrast to expression data, copy number data arises from the projection of discrete values into the experimental space. As such, a reduction of the (a) CGH data can result in the minimal information of segmental status (gain/loss/normal) and genomic position. This facilitates efforts to integrate data across large numbers of experimental series and derived from diverse tumor entities. So far, most of these efforts have been of descriptive nature [10,16] or have been aimed at the definition of disease-specific genomic patterns and useful pattern descriptors ("markers", e.g. [17]). Other publications have attempted the reconstruction of relation and temporal order of oncogenetic events [18-20].

For some cancers types such as subsets of colorectal adenocarcinoma, presence of a limited number of genetic events including several CNA is critical for cancer development [21]. Other neoplasias such as chronic lymphocytic leukemia (CLL) display a paucity of CNA, which however may be correlated to patient survival [22]. These examples illustrate that the presence of certain CNA is not a chance phenomenon, but may either be necessary for cancer development or give a selective edge to affected clones. Previous publications have tried to address the cooperative nature of co-occurring CNA [23,24]. So far, these approaches have not considered the high variability in the complexity of CNAs among individual malignant tumors. Here, we develop an algorithm CDCOCA for analysis of co-occurring oncogenomic CNA events which considers the genomic complexity of the individual samples. We use our approach for detection of CNA events in real-world example data sets. Furthermore, we compare the results from CDCOCA to a previously published method [23] (which we call "analysis 3" in this paper) and also to a modified version of CDCOCA which does not include the adjustment for genomic complexity.

Methods

Data

Annotated copy number and associated data was selected from our Progenetix (a)CGH database ([25]: http://www.progenetix.net; status as of 2010-03-01). For model development and testing, we choose one hematopoietic (MCL) and one solid tumor entity (BLCA) due to their overall intermediate genomic complexity, without consideration of their previously established genomic imbalance profiles or CNA subset analysis.

For analysis, copy number status data was determined for 320 genomic intervals based on corresponding cytogenetic bands. Sex chromosomes were removed due to possible bias in some of the published series (e.g. use as normalization control in (a)CGH experiments), resulting in 303 genomic intervals. For analysis by CDCOCA/CICOCA, gain and loss status of all genomic intervals were considered separately, leading to a data matrix with 606 categories. Only genomic intervals showing change in at least one sample were considered for analysis resulting in a CDCOCA/CICOCA input matrix with 593 categories for BLCA and 571 for MCL. For analysis 3, the original data matrix containing 303 genomic intervals was used. As a surrogate score for genomic complexity, a case specific score was calculated by adding each type of genomic imbalances (gain and/or loss) occurring on a chromosomal arm [26].

From now onwards we will use the term "genomic interval" for genomic interval status. A gain and loss association on same chromosome (e.g. -1p and + 1q) will be referred as "bidirectional" change. The modified structure of the data matrices is exemplified in Table 1. Any gain/loss status of a genomic interval is represented by the value 1.

Model

Let D be the data matrix of dimension nxm, where n is the number of samples and m is the number of genomic intervals. $D_{i,j} = 1$, if a CNA is present in genomic interval j in sample i else $D_{i,j} = 0$. F_j represents the number of sample having CNA at genomic interval j, F_j is given by $\sum_{i=1}^{n} D_{ij}$. $P_w = (P_w^1 \ldots P_w^n)$ represents the vector of probability weights given to samples. The prior probability weight for any sample r *is* defined by the number of CNAs in patient r over total number of CNA across all samples

$$P_w^r = \frac{\sum_{j=1}^{m} D_{rj}}{\sum_{i=0}^{n}\sum_{j=0}^{m} D_{t,j}}$$

Simulation of any genomic interval j is achieved by redistribution of the CNA status over all samples. For genomic interval j, we define $D^{*i} = (D_1^{*1} \ldots D_j^{*n})$ as the corresponding vector representing the CNA status of simulated data. D_j^* is obtained in a way so that $F_j^* \approx F_j$.

Overlay between two genomic intervals is computed using Jaccard's index [27]. Jaccard's index gives a value

Table 1 Binary matrix derived from CGH data

	g-c1p11	g-c1p12	g-c1p13	l-c1p11	l-c1p12	l-c1p13
1	0	0	1	1	1	0
2	0	0	0	0	0	1
3	0	0	1	1	0	0
4	1	1	1	0	0	0
5	1	1	1	0	0	0
6	0	0	0	0	1	1

For each cytogentic band (e.g. c1p12) occurrence of gain (e.g. g-c1p12) and loss (e.g. l-c1p12) is annotated as separate event for each case.

between 0 and 1, where one represents a perfect overlap and zero, no overlap. The Jaccard's index between any two genomic intervals j and k is computed as

$$J_{jk} = \frac{N_{jk}^{11}}{N_{jk}^{10} + N_{jk}^{01} + N_{jk}^{11}}$$

N_{jk}^{11} number of samples with CNA in genomic intervals status, j and k.

N_{jk}^{10} number of samples with CNA in genomic interval status j but not k.

N_{jk}^{01} number of samples with CNA in genomic interval status k but not j.

The overlap obtained on permutation is represented by J_{jk}^{*} Frequency of a co-occurrence is computed as

$$F_{jk} = \frac{N_{jk}^{11}}{n}$$

F_{jk} frequency of an overlap between genomic intervals status i and j.

N_{jk}^{11} number of samples having change in both genomic interval status i and j. n total number of samples in the data.

CDCOCA Algorithm

Let S be the number of simulations and C is the counter measuring the number of times the expected (i.e. permuted) overlap is greater than or equal to the observed overlap. We set the counter of C = 0.

1. Initialize C = 0.
2. Calculate Jaccard's overlap J_{jk} between genomic interval j and k.
3. For genomic interval j simulate the data to obtain D_j^* as
   a. Obtain a sample index r of size 1, from N = (1,....,n) using P_w^i such that sample with maximum weight given has a higher probability of getting a change on permutation, update $D_j^{*r} = 1$.
   b. Update N = N[-r].
   c. Update $P_w^i = P_w^i[-r]$, $P_w^i = \frac{P_w^i}{\sum_i P_w^i}$, $P_w^i = \frac{P_w^i}{1 = P_w^i}$.
   d. Repeat step 3a and 3b F_j times to obtain simulated vector D_j^*.
4. For genomic interval k simulate the data using step 3 to obtain D_k^*.
5. Recompute Jaccard's overlap J_{jk}^*, if $J_{jk}^* \geq j_{jk}$ increase $C = C + 1$.
6. Repeat step 3, 4 and 5 for S times.
7. At the end of S (5000 in our case) permutations calculate p value as, $p = \frac{C}{S}$.

The p-value obtained after step 7 represent the probability of co-occurrence of two CNAs in absence of any other CNA in sample. A low p-value cut off will help in enriching for CNAs which occur together even in less heterogenous samples.

Results and Discussion

We here propose a methodology named CDCOCA (Complexity dependence of co-occurring chromosomal aberrations) that defines highly correlated pairs of CNA in cancer samples while correcting for the overall degree of genomic instability.

We determine CNA complexity based on the number of segmental CNA in a sample while accounting for variations introduced through different resolutions and/or segmentation algorithms [10]. A sample is called "CNA complex" if it has acquired a high number of CNA, and conversely "CNA simple" if a low number of segmental imbalances have been detected. In Figure 1 the distribution of copy number complexities is presented for data from selected tumor entities, extracted from the Progenetix database.

The performance of CDCOCA depends on the number of tumor samples, number of genomic intervals and number of iterations. CDCOCA produces a matrix of p values for all possible associations in the data matrix which are then used to enrich for associations dependent on sample complexity. The algorithm is implemented in the R statistical framework and is available through R package "CDCOCA" provided on the Progenetix website [25].

We applied the CDCOCA algorithm to bladder carcinoma (BLCA) and mantle cell lymphoma (MCL) copy number data, considering gains and losses for each interval as separate events. The readout of the analyses consisted of the p values obtained after randomization for all observed associations in both cancers after 5000 permutations each. We used Jaccard's index to calculate the overlap between genomic intervals [27]. Figure 2 and 3 show the log of p values plotted against the log of Jaccard's index. For simplicity, here p values for only 4 chromosomal changes were plotted. Using CDCOCA we observed that most of the genetic associations have very low Jaccard's overlap and arise

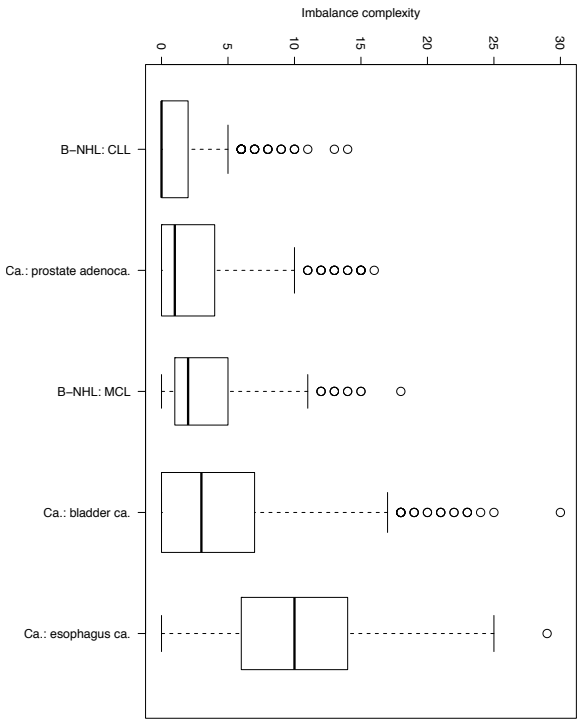


Figure 1 Complexity boxplot of CNA in some selected cancers. Box plot for the overall CNA complexity in selected cancer entities. As a surrogate marker for genomic complexity, each cytogenetic arm was scored independently for gains and losses (i.e., max. score of 4 for a chromosome with both gains and losses on both arms), and chromosome scores were summarized for each case.

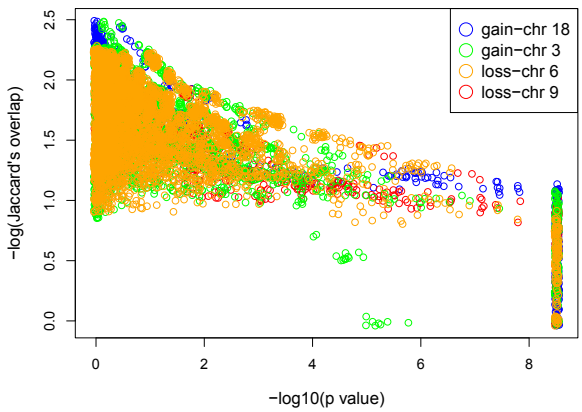


Figure 3 Log of p value plotted against log of Jaccard's index for MCL. Most of the associations have a very high CNA complex background (upper right side of plot) whereas a associations with high Jaccard's index and low p value (lower right side) are also present for all chromosomes.

from genetic changes which occur in CNA complex samples (hence high p values). Associations presenting with high Jaccard's indices and low p-values represent CNA with high probability of specific co-occurrence (i.e. frequent co-occurrence independent of high sample CNA complexity).

Our results show that most of the CNA data for both cancers are derived on a background of multiple and extended CNA. The total number of genetic associations in both cancer types remains beyond scope of the

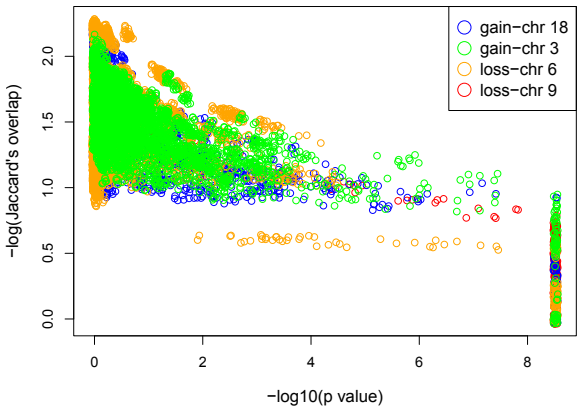


Figure 2 Log of p value plotted against log of Jaccard's index for BLCA. For simplicity reasons all the associations involving only 4 chromosomal changes are shown here. Each color dot reprent an association of that particular chromosome with some other chromosomal band. Most of the associations have a low Jaccard's index and very high p values (upper left side of plot) these associations represent CNA in CNA complex samples. Few associations have a high Jaccard's index and low p values (lower right side of plot); these associations are present in "CNA simple" samples.

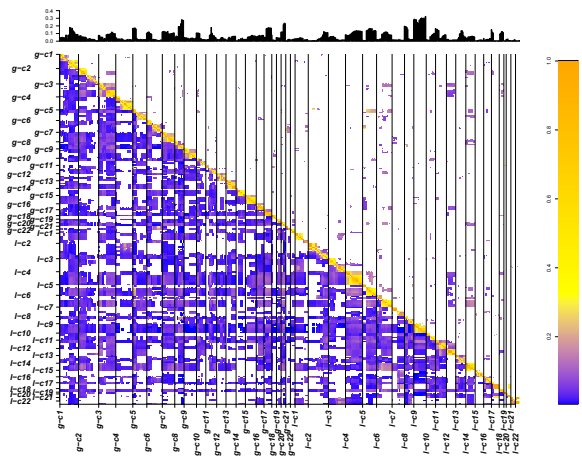


Figure 4 Matrix plot showing results obtained for BLCA with CDCOCA. The diagonal lower half of the image represents all the possible associations in BLCA data where as diagonal upper half represents associations obtained after p value cut off. Frequency of genomic intervals is represented by histogram at the top. Color code represents the value of Jaccard's overlap between associations. The high correlation throughout the diagonal confirms shows the strong connection between CNA co-occurrence and close genetic linkage.

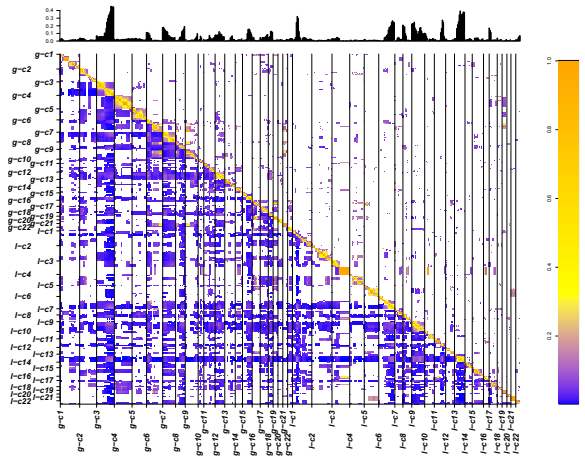


Figure 5 Matrix plot showing results obtained for MCL with CDCOCA. The diagonal lower half of the image represents all the possible associations in BLCA data where as diagonal upper half represents associations obtained after p value cut off. Frequency of genomic intervals is represented by histogram at the top. Color code represents the value of Jaccard's overlap between associations. The high correlation throughout the diagonal confirms shows the strong connection between CNA co-occurrence and close genetic linkage.

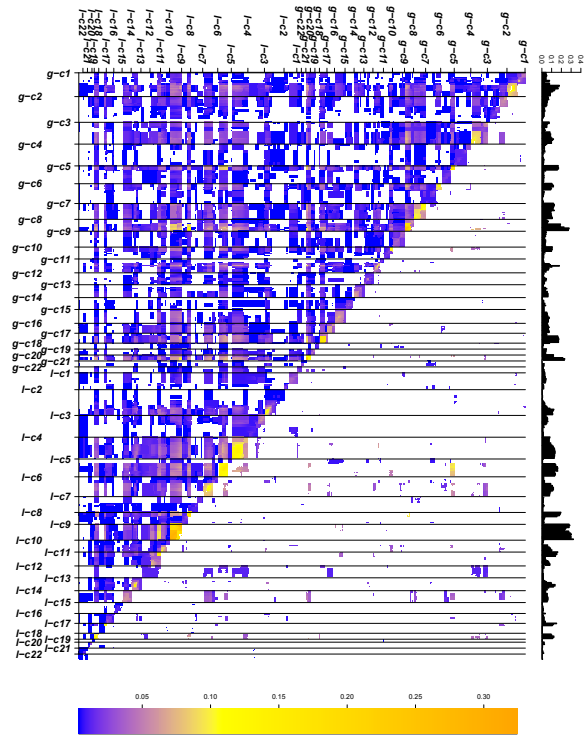


Figure 6 Matrix plot showing frequency of associations in BLCA. Associations are shown similar to figure 4 but here displaying frequencies of associations in place of Jaccard's overlap values. Most frequent CNA associations involve genetically linked CNA. Associations involving high frequent changes are lost after p value cut off indicating their occurrence in a CNA complex background.

current analysis. However, with CDCOCA we are able to focus on a defined set of statistically relevant, specific changes.

For estimating the performance of our methodology in relation to otherwise discussed models we compared CDCOCA to a modified version "CICOCA" (see supplement) and a previously published method [23]. Both the later algorithms do not include an estimate of sample complexity and primarily identify associations with a high frequency. CICOCA and analysis 3 use different methods to compute overlap resulting in slightly different but overall concordant results.

With CICOCA, a high number of co-occurring changes were obtained after p value cut off (Figure 1 and 2 in additional file 1). In contrast, introduction of complexity estimation leads to a focus on changes arising on a low complexity background (Figure 4 and 5). With analysis 3 (Figure 3 and 4 in additional file 1) a very low number of associations was obtained in our sample data set. As expected these only involved high frequent changes. We could show that most of the CNA obtained by analysis 3 (Figure 3 and 4 in additional file 1) were also detected using CDCOCA (Figure 4 and 5) and CICOCA (Figure 1 and 2 in additional file 1). CICOCA and analysis 3 can be used to describe frequent associations, while CDCOCA additionally allows to test the specificity of associations and to apply thresholds accordingly. Compared to frequency based thresholding, one advantage of CDCOCA is its independence from arbitrary cut-off values. The algorithm scores every association. The p value obtained assigns a statistical significance to the associations which is independent of the frequency of the association in the data but takes the complexity of the sample into account.

Table 2 Statistic of associations in BLCA

	Analysis method	Total associations	No. intra-chromosomal associations	p-value	FDR	Associations obtained	No. intera-chromosomal associations
1	CDCOCA	96436	4786	0.02	0.275	6991	3619
2	CICOCA	96436	4786	0.02	0.096	20089	3891
2	Analysis 3	40284	2577	0.02	0.7211	321	152

Bladder carcinoma

An overview of the most frequent genomic imbalances in urinary BLCA can be found in e.g. [10]. Most frequent gains in BLCA include regions on 1q, 5p, 8q,17, 19 and 21q, while the most frequent losses occur on 2q, 4, 5q, 6q, 8p, 9, and 13q (Figure 1 and 3 in additional file 1 and Figure 4 barplot). Due to the high degree of aneuploidy in BLCA, CNA data is highly complex (Figure 4 matrix plot) resulting in a very high number of total associations (Table 2).

A large proportion of associations combine a low frequency with a high Jaccard's index (Figure 4 and 6 matrix plots). We applied a p-value cut off of 0.02 resulting in a false discovery rate (FDR) of 27.5%. At this p-value cut off, 75% of intra-chromosomal associations passed the threshold, confirming the correlation between genetic linkage and involvement in CNA events. Table 2 contains the information about the comparison of results for all three analysis. For simplicity reasons here we limit the display to the 100 most frequent inter-chromosomal changes obtained after p-value cut off.

According to CDCOCA, specific pairs of genomic imbalances in bladder carcinoma include concurrent "bidirectional" losses on 8p and gains on 8q (Figure 7). In the comparative analysis, gains involving chromosome 8q were detected with all three methods (Figure 5 and 6 in additional file 1 and Figure 7). However, with CDCOCA the frequent co-occurrence of these CNA on the background of a low genomic complexity became more apparent. This observation may point to an early appearance of these CNA during tumorigenesis, with a possible role as cancer initiating event. While gains on distal 8q are the most consistent copy number change in epithelial neoplasias with MYC considered a predominant target, recently deletions on 8p23.3 have been associated with aggressive clinical behavior in BLCA [28]. Another observation concerned changes involving concurrent gains on 5p and losses on 5q which were also associated with losses on chromosome 4q and distal 6q (6q22). These co-occurrences (Figure 5 and 6 in additional file 1 and Figure 7). Although one may assume that "bidirectional" changes involving both chromosomal arms are based on simple cytogenetic events, e.g. isochromosome formation, the limitation of this pattern to distinct chromosomes points at an evolutionary advantage of both gain and

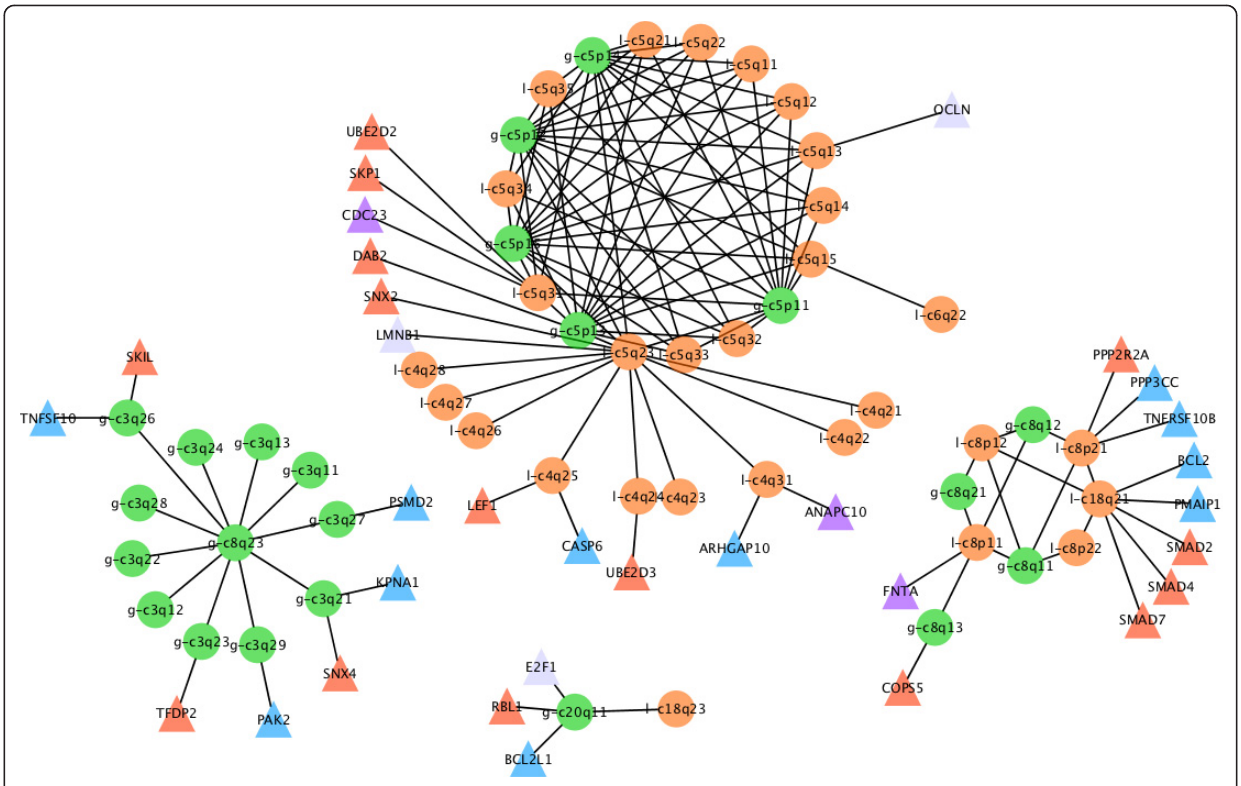


Figure 7 100 frequent associations obtained after p value cut off in BLCA with CDCOCA. Pathway associations for presumptive target genes from the 100 most frequent associations in BLCA obtained after p value cut-off are shown. TGF-beta signaling genes (blue triangles) and genes from cellular apoptotic pathways (red triangles) present in these genetic locations are displayed. Gains are represented with prefix "g" and green color circles whereas losses with prefix "l" and orange circles. Associations were drawn using cytoscape [42].

loss accumulation for the malignant clone. Other event pairs obtained by CDCCOA include gains on 8q23 along with gains on 3q, as well as gain on 20q11 with loss on 18q23.

The abundance of 8p losses, 8q gains, 5q losses, 5p gains, 3q gains, 4q losses points towards the importance of these CNA in tumors carrying them. Genes from TGF-beta receptor signaling (blue triangles) and cellular apoptotic pathways (red triangles) located to the co-occurring changes are shown in Figure 7. The presence of genes from the same pathways on co-occurring CNA point towards a possible cooperative action of these genes. CDC23 (5q31), CASP6 (4q25) and PMAIP1 (18q21) are among TGF-receptor cascade genes with well established role in cancer [29,30] Other possible targets for genetic cooperation include PMSD2, PAK2, BCL2L1 and FNTA. Genes from apoptotic signaling pathways mapped to these regions include CDC23 (5q31), SMAD2 (18q21), SMAD4 (18q21) and SMAD7 (18q21) which have been shown defective in several cancer entities [31]. As possible target on 5p, loss of SKP2 had been shown to cause cell senescence [32]. On 5q, loss of function mutations including copy number losses of both APC and MCC have been associated with a variety of epithelial neoplasias [33-36].

Mantle cell lymphoma

For MCL, an overall p value distribution similar to that of BLCA was observed (Figure 3). Most common CNA in MCL included gains on chromosomes 3q, 6p, 7p and 8q, while most common losses involved regions on 6q, 8p, 9, 11q and 13q (Figure 7 and 8 in additional file 1 and Figure 5).

A p value cut-off of 0.04 giving a FDR of 30% was applied with CDCOCA (Table 3 and Figure 8). About 80% of intra-chromosomal associations passed this threshold, representing approx. 50% of all post cut-off associations. The 100 strongest associations obtained with CDCOCA are shown in Figure 9. As in BLCA, CDCOCA detected losses on 8p with gains on 8q, which was not described as association in the other analyses. Also, only CDCOCA selected groups of co-occurrences involving low frequency CNA (e.g. associations involving gains 7p, 6p, 12p and 18q). Other changes such as losses on highly occurring 13q along with gains on not so frequently occurring 7q were obtained using CDCOCA and not with CICOCA and analysis 3 in the top 100 events (Figure 7 and 8 in additional file 1).

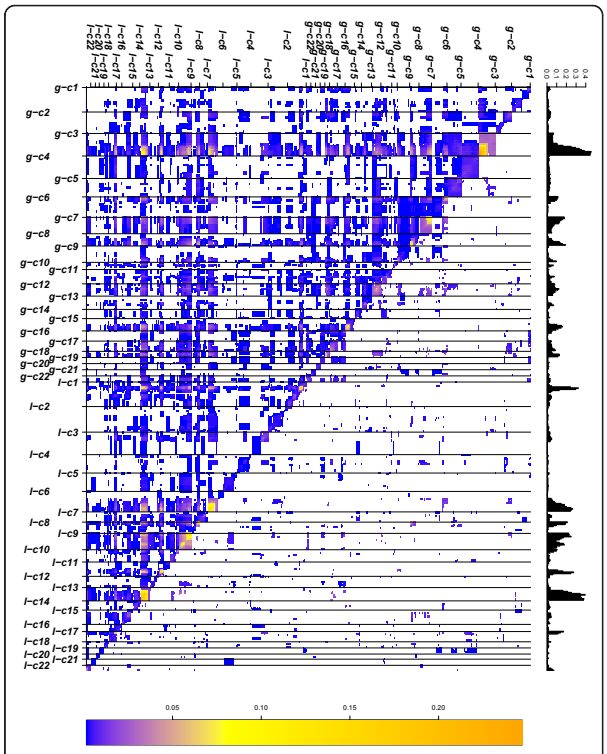


Figure 8 Matrix plot showing frequency of associations in MCL. Associations are shown similar to figure 7 but here displaying frequencies of associations in place of Jaccard's overlap values. Most frequent CNA associations involve genetically linked CNA. Associations involving high frequent changes are lost after p value cut off indicating their occurrence in a CNA complex background.

As candidate targets, TNF-signaling genes (red triangles) and T-cell receptor signaling genes (blue triangles) are marked on their corresponding band locations in Figure 9. The role of genes such as MDM2 (12p15), TNFRSF1A (12p13), MALT1 (18q21) for neoplastic transformation and/or progression has already been well established [37-39]. Other examples for cancer relevant genes mapping to those regions are STAT2 (12q13), and STAT3 (17q) [40,41].

Conclusions

We have developed a method CDCOCA to define complexity dependence of co-occurring CNA in cancer

Table 3 Statistic of associations in MCL

	Analysis method	Total associations	No. intra-chromosomal associations	p-value	FDR	Associations obtained	No. intera-chromosomal associations
1	CDCOCA	57644	3918	0.04	0.30	7513	3175
2	CICOCA	57644	3918	0.04	0.197	11673	3918
2	Analysis 3	31136	2418	0.04	0.571	867	207

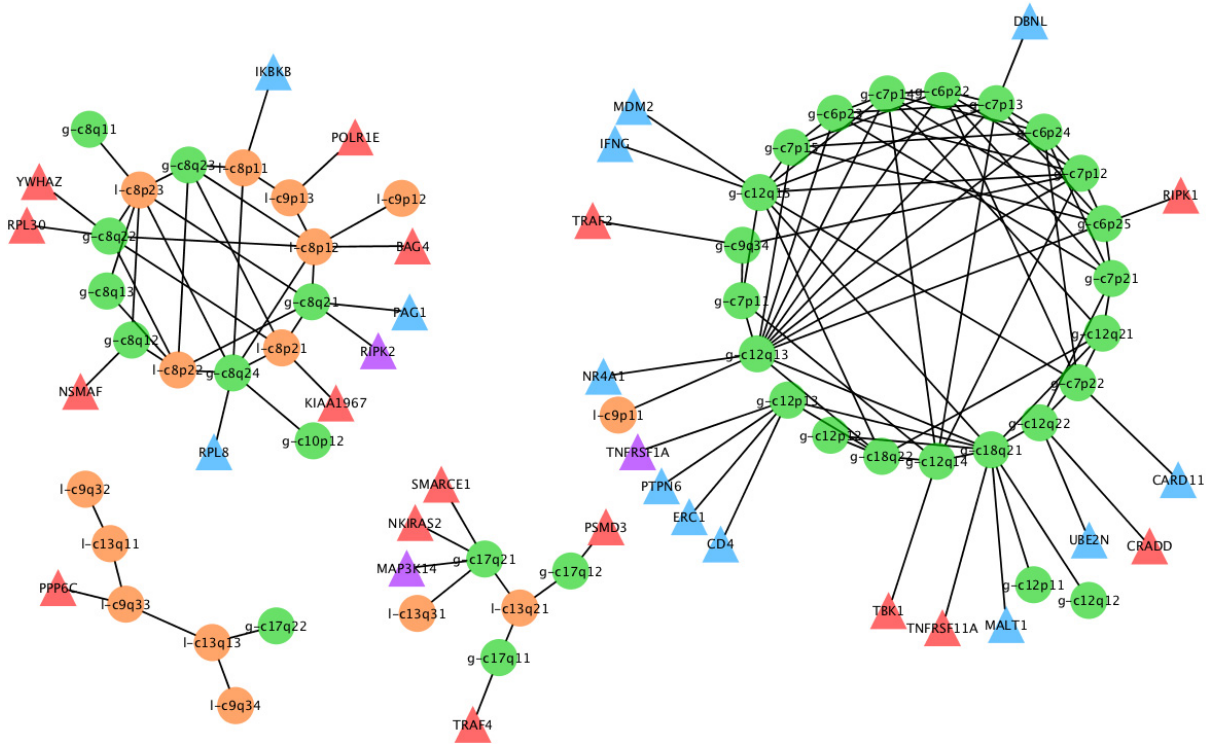


Figure 9 100 frequent associations obtained after p value cut off in MCL with CDCOCA. Pathway associations for presumptive target genes from the 100 most frequent associations in MCL. TNF-alpha signaling genes (red triangles) and T-cell receptor signaling pathway genes (blue triangles) present in these genetic locations are shown. Gains are represented with prefix "g" and green color circles whereas losses with prefix "l" and orange circles.

samples. In contrast to methods published previously [23] and a modified algorithm which does not include the complexity adjustment step, CDCOCA does not simply focus on the most frequent co-occurrences of regional genomic copy number changes in cancer entities. Here, we determine statistically relevant co-occurring CNA through accounting for the CNA "background noise", introduced e.g. through chromosome scale imbalances (e.g. isochromosomes, chromosomal aneuploidy). In theory, this procedure should highlight specific but comparatively rare CNA events.

As indicated by our analysis of BLCA and MCL, two unrelated cancer entities with overall intermediate copy number complexity, the relevant CNA associations in many specimen are obscured due to the large number and/or extension of regional CNA. When correcting for genomic background heterogeneity most of the associations involving highly recurring CNA were removed. This indicates that many high frequency changes may be related to the overall genomic instability and therefore cannot unanimously be assigned a causative role in oncogenesis. Especially regarding the large number of genes affected by complex genomic imbalances, some of the cancer type specific CNA patterns may represent an epiphenomenon of disturbed genomic maintenance processes rather than the expression of copy number dependent target gene modifications.

However, when accounting for the overall complexity, CNA associations may point towards connected events and/or preferred pathways activated during carcinogenesis. Based on our CNA associations, we found multiple genes from single well defined cancer pathways to be a effected in sample subsets. Alteration of more than one gene in a pathway may potentiate the effect on pathway function and be responsible for a specific clonal phenotype.

CDCOCA should prove to be a powerful tool for defining mutual associations at gene level and to gain insights into cellular mechanisms relevant for oncogenesis. Although we applied our method to CGH data at band resolution, there is no practical obstacle against use with segmented data from high resolution genomic array experiments. In fact, this should facilitate a gene centric analysis and automatic integration with functional data sources.

Additional material

Additional file 1: CICOCA: A method to define complexity independence of co-occurring chromosomal aberrations. The additional file contains information about the statistical method CICOCA which is compared with CDCOCA. This method (as described in text above) aims in finding co-occurring chromosomal associations independent of the sample complexity. In addition to CICOCA this file also contains all the additional figures which are referred in the paper along with a detail description of all the additional figures.

Acknowledgements
NK is supported through a grant by the Krebsliga Schweiz (Swiss Cancer League). Haoyang Cai is supported through a grant from the China Scholarship Council.

Author details
[1]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich, Switzerland. [2]Functional Genomics Center Zurich, University of Zurich, Winterthurerstrasse 190, Zurich, Switzerland.

Authors' contributions
NK, MB, HR designed and conceived the experiments; NK implemented the software; NK, MB analyzed the data, HC, MB contributed to the data. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 26 July 2010 Accepted: 3 March 2011
Published: 3 March 2011

References

1. Futreal P, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton M: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4(3)**:177-83.
2. Stratton M, Campbell P, Futreal P: **The cancer genome.** *Nature* 2009, **458(7239)**:719-24.
3. Kinzler K, Vogelstein B: **Lessons from hereditary colorectal cancer.** *Cell* 1996, **87(2)**:159-70.
4. Hanahan D, Weinberg R: **The hallmarks of cancer.** *Cell* 2000, **100**:57-70.
5. Lengauer C, Kinzler K, Vogelstein B: **Genetic instabilities in human cancers.** *Nature* 1998, **396(6712)**:643-9.
6. Stallings R: **Origin and functional significance of large-scale chromosomal imbalances in neuroblastoma.** *Cytogenet Genome Res* 2007, **118(2-4)**:110-5, [Copyright (c) 2007 S. Karger AG, Basel.].
7. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirieac LR, Cho J, Fujii Y, Gazdar AF, Giordano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, Mardis ER, McPherson JD, Minna JD, Morgan MB, Nadel M, Orringer MB, Osborne JR, Ozenberger B, Ramos AH, Robinson J, Roth JA, Rusch V, Sasaki H, Shepherd F, Sougnez C, Spitz MR, Tsao MS, Twomey D, Verhaak RGW, Weinstock GM, Wheeler DA, Winckler W, Yoshizawa A, Yu S, Zakowski MF, Zhang Q, Beer DG, Wistuba II, Watson MA, Garraway LA, Ladanyi M, Travis WD, Pao W, Rubin MA, Gabriel SB, Gibbs RA, Varmus HE, Wilson RK, Lander ES, Meyerson M: **Characterizing the cancer genome in lung adenocarcinoma.** *Nature* 2007, **450(7171)**:893-8.
8. Myllykangas S, Himberg J, Bohling T, Nagy B, Hollmen J, Knuutila S: **DNA copy number amplification profiling of human neoplasms.** *Oncogene* 2006, **25(55)**:7324-32.
9. Coe B, Lockwood W, Girard L, Chari R, Macaulay C, Lam S, Gazdar A, Minna J, Lam W: **Di erential disruption of cell cycle pathways in small cell and non-small cell lung cancer.** *Br J Cancer* 2006, **94(12)**:1927-35.
10. Baudis M: **Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data.** *BMC Cancer* 2007, **7**:226.
11. Kallioniemi A, Kallioniemi O, Sudar D, Rutovitz D, Gray J, Waldman F, Pinkel D: **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.** *Science* 1992, **258(5083)**:818-21.
12. du Manoir S, Speicher M, Joos S, Schrock E, Popp S, Dohner H, Kovacs G, Robert-Nicoud M, Lichter P, Cremer T: **Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization.** *Hum Genet* 1993, **90(6)**:590-610.
13. Joos S, Scherthan H, Speicher M, Schlegel J, Cremer T, Lichter P: **Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe.** *Hum Genet* 1993, **90(6)**:584-9.
14. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P: **Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances.** *Genes Chromosomes Cancer* 1997, **20(4)**:399-407.
15. Pinkel D, Albertson D: **Comparative genomic hybridization.** *Annu Rev Genomics Hum Genet* 2005, **6**:331-54.
16. Beroukhim R, Mermel C, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm J, Dobson J, Urashima M, Henry KM, Pinchback R, Ligon A, Cho Y, Haery L, Greulich H, Reich M, Winckler W, Lawrence M, Weir B, Tanaka K, Chiang D, Bass A, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye F, Sasaki H, Tepper J, Fletcher J, Tabernero J, Baselga J, Tsao M, Demichelis F, Rubin M, Janne P, Daly M, Nucera C, Levine R, Ebert B, Gabriel S, Rustgi A, Antonescu C, Ladanyi M, Letai A, Garraway L, Loda M, Beer D, True L, Okamoto A, Pomeroy S, Singer S, Golub T, Lander E, Getz G, Sellers W, Meyerson M: **The landscape of somatic copy-number alteration across human cancers.** *Nature* 2010, **463(7283)**:899-905.
17. Liu J, Ranka S, Kahveci T: **Markers improve clustering of CGH data.** *Bioinformatics* 2007, **23(4)**:450-7.
18. Hoglund M, Frigyesi A, Sall T, Gisselsson D, Mitelman F: **Statistical behavior of complex cancer karyotypes.** *Genes Chromosomes Cancer* 2005, **42(4)**:327-41, [(c) 2005 Wiley-Liss, Inc.].
19. Desper R, Jiang F, Kallioniemi O, Moch H, Papadimitriou C, Schaffer A: **Distance-based reconstruction of tree models for oncogenesis.** *J Comput Biol* 2000, **7(6)**:789-803.
20. Gerstung M, Baudis M, Moch H, Beerenwinkel N: **Quantifying cancer progression with conjunctive Bayesian networks.** *Bioinformatics* 2009, **25(21)**:2809-15.
21. Vogelstein B, Fearon E, Hamilton S, Kern S, Preisinger A, Leppert M, Nakamura Y, White R, Smits A, Bos J: **Genetic alterations during colorectal-tumor development.** *N Engl J Med* 1988, **319(9)**:525-32.
22. Dohner H, Stilgenbauer S, Benner A, Leupolt E, Krober A, Bullinger L, Dohner K, Bentz M, Lichter P: **Genomic aberrations and survival in chronic lymphocytic leukemia.** *N Engl J Med* 2000, **343(26)**:1910-6.
23. Bredel M, Scholtens D, Harsh G, Bredel C, Chandler J, Renfrow J, Yadav A, Vogel H, Scheck A, Tibshirani R, Sikic B: **A network model of a cooperative genetic landscape in brain tumors.** *JAMA* 2009, **302(3)**:261-75.
24. Klijn C, Bot J, Adams D, Reinders M, Wessels L, Jonkers J: **Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach.** *PLoS Comput Biol* 2010, **6**: e1000631.
25. Baudis M, Cleary ML: **Progenetix.net: an online repository for molecular cytogenetic aberration data.** *Bioinformatics* 2001, **17(12)**:1228-9.
26. Boerma E, Siebert R, Kluin P, Baudis M: **Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of todays knowledge.** *Leukemia* 2009, **23**:225-234.
27. Tan PN, Steinbach M, Kumar V: *Introduction to data mining* Bosotn, MA, USA: Addison Wesley; 2005.
28. Eguchi S, Yamamoto Y, Sakano S, Chochi Y, Nakao M, Kawauchi S, Furuya T, Oga A, Matsuyama H, Sasaki K: **The loss of 8p23.3 is a novel marker for predicting progression and recurrence of bladder tumors without muscle invasion.** *Cancer Genet Cytogenet* 2010, **200**:16-22, [2010 Elsevier Inc. All rights reserved.].
29. Wang Q, Moyret-Lalle C, Couzon F, Surbiguet-Clippe C, Saurin J, Lorca T, Navarro C, Puisieux A: **Alterations of anaphase-promoting complex genes in human colon cancer cells.** *Oncogene* 2003, **22(10)**:1486-90.
30. Loro L, Johannessen A, Vintermyr O: **Loss of BCL-2 in the progression of oral cancer is not attributable to mutations.** *J Clin Pathol* 2005, **58(11)**:1157-62.

31. Maliekal T, Antony M, Nair A, Paulmurugan R, Karunagaran D: **Loss of expression, and mutations of Smad 2 and Smad 4 in human cervical cancer.** *Oncogene* 2003, **22(31)**:4889-97.
32. Lin HK, Chen Z, Wang G, Nardella C, Lee SW, Chan CH, Yang WL, Wang J, Egia A, Nakayama KI, Cordon-Cardo C, Teruya-Feldstein J, Pandolfi PP: **Skp2 targeting suppresses tumorigenesis by Arf-p53-independent cellular senescence.** *Nature* 2010, **464(7287)**:374-9.
33. Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M, al et: **Identification and characterization of the familial adenomatous polyposis coli gene.** *Cell* 1991, **66(3)**:589-600.
34. Kinzler K, Nilbert M, Vogelstein B, Bryan T, Levy D, Smith K, Preisinger A, Hamilton S, Hedge P, Markham A, al et: **Identification of a gene located at chromosome 5q21 that is mutated in colorectal cancers.** *Science* 1991, **251(4999)**:1366-70.
35. Nishisho I, Nakamura Y, Miyoshi Y, Miki Y, Ando H, Horii A, Koyama K, Utsunomiya J, Baba S, Hedge P: **Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients.** *Science* 1991, **253(5020)**:665-9.
36. Ashton-Rickardt P, Wyllie A, Bird C, Dunlop M, Steel C, Morris R, Piris J, Romanowski P, Wood R, White R, al et: **MCC, a candidate familial polyposis gene in 5q.21, shows frequent allele loss in colorectal and lung cancer.** *Oncogene* 1991, **6(10)**:1881-6.
37. Trauzold A, Roder C, Sipos B, Karsten K, Arlt A, Jiang P, Martin-Subero J, Siegmund D, Muerkoster S, Pagerols-Raluy L, Siebert R, Wajant H, Kalthoff H: **CD95 and TRAF2 promote invasiveness of pancreatic cancer cells.** *FASEB J* 2005, **19(6)**:620-2.
38. Sugano N, Suda T, Godai T, Tsuchida K, Shiozawa M, Sekiguchi H, Yoshihara M, Matsukuma S, Sakuma Y, Tsuchiya E, Kameda Y, Akaike M, Miyagi Y: **MDM2 gene amplification in colorectal cancer is associated with disease progression at the primary site, but inversely correlated with distant metastasis.** *Genes Chromosomes Cancer* 2010, **49(7)**:620-9, [(c) 2010 Wiley-Liss, Inc.].
39. Dierlamm J, Penas EM, Bentink S, Wessendorf S, Berger H, Hummel M, Klapper W, Lenze D, Rosenwald A, Haralambieva E, Ott G, Cogliatti S, Moller P, Schwaenen C, Stein H, Loffer M, Spang R, Trumper L, Siebert R: **Gain of chromosome region 18q21 including the MALT1 gene is associated with the activated B-cell-like gene expression subtype and increased BCL2 gene dosage and protein expression in diffuse large B-cell lymphoma.** *Haematologica* 2008, **93(5)**:688-96.
40. Konnikova L, Simeone M, Kruger M, Kotecki M, Cochran B: **Signal transducer and activator of transcription 3 (STAT3) regulates human telomerase reverse transcriptase (hTERT) expression in human cancer and primary cells.** *Cancer Res* 2005, **65(15)**:6516-20.
41. He B, Reguart N, You L, Mazieres J, Xu Z, Lee A, Mikami I, McCormick F, Jablons D: **Blockade of Wnt-1 signaling induces apoptosis in human colorectal cancer cells containing downstream mutations.** *Oncogene* 2005, **24(18)**:3054-8.
42. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13**:2498-2504.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1755-8794/4/21/prepub

doi:10.1186/1755-8794-4-21
Cite this article as: Kumar *et al.*: CDCOCA: A statistical method to define complexity dependence of co-occuring chromosomal aberrations. *BMC Medical Genomics* 2011 **4**:21.

9.2 Publication 2: Signaling pathway enrichment in cancer copy number alteration data (Manuscript in preparation)

Signaling pathway enrichment in cancer copy number alteration data

Nitin Kumar[1], Hubert Rehrauer[+2] , Sushil Kumar[+3] , Haoyang Cai [1] and Michael Baudis[*1]

[1]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich, Switzerland,
[2]Functional Genomics Center Zurich, University of Zurich, Winterthurerstrasse 190, Zurich, Switzerland,
[3]Institute of Molecular Cancer Research, University of Zurich, Winterthurerstrasse 190, 8046 Zurich Switzerland
[4]Swiss Institute of Bioinformatics, Quartier Sorge, Lausanne, Switzerland
[+]equal contribution

Email: Nitin Kumar - nitin.kumar@imls.uzh.ch; Hubert Rehrauer - hubert.rehrauer@fgcz.ethz.ch; Sushil Kumar - kumar@imcr.uzh.ch; Haoyang Cai - haoyang.cai@imls.uzh.ch; Michael Baudis *- michael.baudis@imls.uzh.ch;

[*]Corresponding author

Abstract

Regional genomic copy number aberrations (CNA) are a type of mutation observed in the vast majority of cancer entities. Since many of the CNA in cancer may simply represent neutral somatic variants, and since they typically cover several genes simultaneously, their causal contribution to tumorigenesis has been difficult to quantify. The detection of systematically enriched functional pathways among CAN-affected genes would be one way to assess causality and such an observation would also be instrumental in better describing cancer pathways with respect to their entire spectrum of mutations.

We here use three different strategies (H-path [1], G-path and S-path) to search for pathways affected by CNA more often than expected by chance. In addition to pathway enrichment S-path also considers the key issue of genomic clustering of cellular pathways. The analysis is performed on CNA data from a total of 19819 tumor samples. The new strategy S-path outperforms the other two methods, but still finds only a very restricted set of pathways that might be somewhat enriched. The analysis is then extended to 132 individual cancer types classified on the basis of international classification of disease code (ICD) to define a cancer specific pathway signature. We here show that some pathways are altered by CNA more often than chance and are enriched among multiple cancer types. A knowledge driven analysis of enriched pathways lead to the definition of specific functions controlling cancer associated cellular responses.

Background

Somatic mutations are an essential requirement for the development of a neoplastic cell clone. Regional deviations of the DNA copy number from the normal diploid status (termed "copy number alterations/aberrations", CNA) can be observed in the majority of malignant neoplasias. Several regional CNA hot-spots (i.e. local peaks of recurring gains / losses) have been linked to specific oncogenes and tumor suppressor genes such as REL (2p13-12), p53 (17p13), MYCN (2p24). Some recurring CNA are present across multiple cancer types [2, 3]. However, the extensive CNA heterogeneity observed in human malignancies may obscure any underlying activation of common growth mechanisms, because different components of tumor-related pathways will be affected in each patient [4, 5].The determination of cellular pathways systematically enriched in CNA regions would provide insights in the relative importance of functional mechanisms for the development of specific neoplastic entities, and should lead to an extensive delineation of specific CNA targets.

Previously, extensive pathway analyses have been performed using expression array data [6–11]. Additionally, some studies have reported integrative analyses of CNA and expression array data to predict pathways altered in cancers [12, 13], or have combined CNA with somatic mutation data [1]. Methods have also been developed to detect enriched pathways in cancer SNP data [14, 15]. However, to the best of our knowledge no analysis has been performed to search for enriched cellular pathways solely based on cancer CNA data across multiple cancer types. Such an unbiased analysis should prove especially useful due to the potential for analysis across very large CNA datasets available through public resources.

Here, we first propose two new permutation methodology, "G-path" (gene based pathway enrichment) and "S-path" (for segment based pathway enrichment), to detect enriched cellular pathways among the genomic regions affected by copy number changes in cancer. S-path is a variant of G-path, which defines pathways as a set of genomic segments and not genes (see Methods); this considers the key issue of genomic clustering of cellular pathway genes which may otherwise lead to false-positive calls of pathway enrichments [16]. The analysis is also compared to a know pathway enrichment methodology [1] (supplement, referred to as "H-path" for pathway hit enrichment). Both G-path and S-path consider the following two points about pathway-based analysis in CNA data:

- A cellular pathway consists of multiple interacting proteins, modulators and e?ectors, and hence multiple genes/segments may be simultaneously altered in any individual tumor, potentially conferring a certain growth-advantage.
- Annotated pathways may often partially overlap, and also cover widely different levels of functional specificity, so both gene-pathway and segment pathway (for S-path) membership are conserved while performing a whole genome permutation strategy, in order to search for statistically enriched pathways.

We first tested the three algorithms on a cancer dataset encompassing 19819 tumor samples available through our website Progenetix [17] and chose the best performing algorithm on the basis of the number of pathways reported as enriched below a given FDR threshold. The analysis was then extended to 132 individual cancer types classified with the help of international classification of disease (ICD) codes. High-scoring pathways were than inspected to determine how they may interact and what kind of cellular response they might generate.

Methods

Gene and pathway data

Several resources of pathway-associated annotations are publicly available (e.g. Reactome [18], Kegg [19], Nature Signaling Pathways [20]). We use Nature Signaling Pathways as a pathways source for all analyses reported here; it consists of signaling cascades only and does not include processes such as metabolism (present in KEGG). Another reason for using Nature Signaling Pathways is its non-hierarchal pattern of annotation, contrary to e.g. Reactome where large pathways are further divided into hierarchies of pathways making analysis difficult. Genes were mapped to genomic locations according to Human genome version 18 (hg18/build36). Genes on chromosomes X and Y were removed prior to analysis due to inherent gender biases (e.g. prostate carcinoma) and inconsistent reporting in the CNA data sets. The remaining pathway-specific data consisted of 176 pathways containing a total of 2239 genes.

For genomic positions, we obtained the Ensembl gene list from Biomart release 54. This gene list was processed to obtain genes with unique combination of Ensembl gene identifier (Ensembl id) and chromosome start and end position. Mitochondrial and sex chromosomal genes were removed resulting in a total of 20209 genes for the input gene list.

For S-path we divided all chromosomes in to non-overlapping segments of 1Mb each, dividing the entire genome into 2872 bins (we call this "segmentation"). The gene list was than mapped to the artifical

segments of 1Mb, and each segment containing one or more genes of a given pathway was counted exactly once for that pathway. We expected that dividing the genome into small bins resolved the issue of pathway genes clustering. 2239 pathway genes were mapped to 1203 unique genomic segments.

Data sets

Cancer CNA data available through progenetix [17] is used as an input to pathway algorithms. A total of 19819 tumor samples from 132 different cancer types were analyzed for pathway enrichment. The analysis was then extended to individual cancer type for entity specific pathway enrichment profile. The genome for copy number data across all samples was reduced to segments of 1Mb each as an input to S-path. Such an input data has been used to enrich for CNA which are altered more often than chance [21].

Model and parameters

We here define models and parameters for G-path and S-path. The only difference between G-path and S-path is the pathway input file. Genes are an input for G-path which are replaced with segments for S-path. Here we use the term genes for both; genes for G-path and segments for S-path.
A unique permutation strategy is used to compute pathway scores using G-path/S-path. Let S represents a list of samples $\{S_1, ...S_n, ...S_m\}$ of a given cancer type. A pathway P is represented as a set of genes G_i , where i is any number of genes with known genomic locations on autosomes. We consider that for pathway P more than one gene can simultaneously be affected by CNA Figure 1 shows the boxplot of number of 1Mb pathway segments altered in some pathways across 19379 tumor samples, showing on average 10% of segments are copy number altered across pathways. For every sample a pathway index $SP_n = N_P$ is obtained. N_P represents number of genes for pathway P which are copy number altered in sample S_n, N_P ϵ G_i. The overall pathway score for pathway P across all samples is obtained as $OP_P = \sum_{n=1}^{m} SP_n$. The list of overall pathway scores is represented as $\{OP_1, ...OPp, ...OPs\}$for s pathways. The corresponding list obtained on permutations is represented as $\{OP_1^*, ...OPp^*, ...OPs^*\}$. OP_p can vary between 0 (when no gene for pathway P is copy number altered) and $\max\{i * m\}$(when all genes i are copy number altered across all m samples). All the three algorithms (G-path, S-path and H-path) identify pathways rejecting the null hypothesis that "there is no association between pathway genes and CNA".

Pathways enrichment with G-path and S-path

In this section we describe the two algorithm G-path and S-path. Genes for G-path represent same as segments for S-path.

1. Overall pathway scores for all pathways $\{OP_1, ...OP_p, ...OP_s\}$are computed.
2. A null vector $C = \text{NULL}$ of length s is defined. C measures the index how many times expected scores across pathways are greater than or equal to observed scores.
3. For permutation all Ensemble genes (2872 segments for S-path) are randomly distributed on to new genomic locations. This permutation helps in keeping gene pathway membership consistent while affecting only pathways Vs genomic location membership.
4. Expected pathway score $\{OP_1^*, ...OP_p^*, ...OP_s^*\}$on permutation for all pathways is computed.
5. If expected score for pathway P is greater than or equal to the observed score C_p is incremented as

$$C_p = \left\{ \begin{array}{ll} C_p + 1 & if\ OP_p^* >= OP_p \\ C_r & if\ OP_p^* < OP_p \end{array} \right\}$$

6. Step 2 to 5 are repeated over Z number of times (Z = 10000 in current analysis) increasing the counter C for pathways which are altered by chance.
7. After all permutations a p value vector $\{p_1, ..p_p, ..p_s\}\ \epsilon\ P$ are calculated from C as $p_p = \frac{C_p}{Z}$
8. All the p values are corrected for false discovery rate using Benjamini Hochberg correction.

Genomic clustering of pathways

Genes for cellular pathways are clustered on the genome [16]. For all the pathways a clustering score is compute as described in [16]. Since for S-path genes from pathways are represented as segments the pathway clustering score is computed using these segments and not genes. For each gene pair on same chromosome from a pathway a pair wise score of clustering is computed as

$$pairscore = \frac{average\ length\ of\ chromosomes\ in\ genome}{distance\ between\ genes}$$

For genes on different chromosome distance is computed as

$$pairscore = \frac{average\ length\ of\ chromosomes\ in\ genome}{average\ length\ of\ chromosomes\ the\ genes\ are\ located\ on}$$

The pathway clustering score is the sum of all pair wise scores divided by the number of genes in that pathway. The original clustering score is compared to the expected scores obtained by randomly creating pathways with same number of genes and then recalculating the expected pathway clustering score (10000

permutations). p-values are generated by comparing the original and expected scores as $pvalue = \frac{number\ of\ times\ expected\ clustering\ score \geq observer\ clustering\ score}{number\ of\ permutations}$.

The p values are corrected for false discovery rate using Benjamini Hochberg correction.

Results and Discussion

Clustering of pathway genes

The pathway clustering score followed nearly a normal distribution with data having a mean score of 89424.79 (Figure 2). The clustering score was independent of the number of genes in pathways (Supplement figure 1). Dividing the genome into artificial segments of 1Mb resulted in a suppression of genomic clustering (Figure 2) and the mean clustering score was reduced to 114.465. Genomic pathway clustering signal was lost after genome segmentation (supplement figure 1). Before segmentation 47 pathways were found to be genomic clustered at this FDR (Supplement table 1). However when segments were considered none of the pathways remained clustered (supplement figure 1). This analysis points towards a key issues about considering these genomic segments as a set of pathway components and not genes. As considering genes produce a signal which is affected by the genomic clustering of pathways.

Comparison H-path, G-path and S-path

We compared all the three methods to obtain a list of pathways targeted by can more often than expected by chance. Cancer CNA alterations span through multiple pathway genes (Figure 1), indicating the importance of the total score as a pathway score parameter. A maximum FDR cutoff of 0.15 was used as a selection criteria to find enriched pathways. H-path led to an enrichment of only one pathway (Supplement figure 2) whereas 4 pathways (one genomic unclustered and three clustered) were enriched with G-path (Supplement figure 3). Using S-path 17 pathways were found to be significantly enriched (Figure 3). Genes from five of these pathways were clustered in genome however since we considered segments this genomic clustering was not significant and did not affect our results.

All the pathways enriched in G-path were enriched using S-path however additional pathways were identified using S-path with same FDR cut-offs. The issue of genomic clustering is also resolved by S-path as all the genes present in a window bin of 1Mb are considered as a single event. H-path efficiently identifies pathways which are hit more often with CNA than by chance. However the analysis lacks the ideology that for a given pathway more than one gene can be simultaneously copy number altered. The permutation strategy used by G-path and S-path is unique in itself as till now no one has considered to

keep genes (segment) and pathway membership constant on permutations. Another advantage of G-path/S-path is that a much large range of background CNA are considered in analysis.

Pathways enriched

We extended the analysis using S-path to CNA data from 132 different cancer types. 59 cancer types showed an enrichment of at least one pathway with a FDR cut off of 0.15. 22 pathways were enriched in more than 20 individual cancer types (Supplement figure 4). On further scrutiny we nailed down to 8 signaling pathways (Figure 4), which were identified in integrative analysis as well as were present in more than 20 ICD types. Interesting candidates included mTOR signaling pathway and Thromboxane A2 receptor signaling, both of these pathways were enriched in more than 30 cancer types.

Identification of these pathways predicts increased proliferation or antiapoptotic properties in cancer cell (Figure 4). For instance, although a6b integrin are known for their function in cellular adhesion to the extracellular matrix, recent studies have added a new oncogenic role of these integrins through enhancement of erbB and MET signaling pathways [22, 23]. In the samples we analyzed, a6b integrin gene loci are gained concurrently with erbB receptors and MET (Supplement Table 1), suggesting a potentiation of erbB/MET signaling through cooperation with these integrins. The PTP1b pathways that are negative regulators of receptor tyrosine kinase (RTK) signaling also showed gain for genes encoding growth factor receptors (Supplement Table 1) or their downstream positive regulators [24–26]. Other important pathways that are found significantly modulated in terms of gene copy number include thromboxane signaling pathway which is a GPCR signaling pathway also known to activate growth factor signaling by a novel mechanism known as receptor transactivation.

The mechanism of cellular proliferation is dissected into cyclic entry of cells in mitosis regulated by set of transcription factors. An important class of such transcription factors is E2F family. Several member of this family were found increased in copy number. Interestingly, an important tumor suppressor called pRb is lost in our samples, which strengthen the confidence in our algorithm in identifying important signaling pathways, and their genes involved in tumor development. The mammalian target of rapamycin (mTOR) is known as growth regulator of cells by its virtue of regulating protein synthesis, metabolism, and autophagy in response to converging signals transduced from cell surface receptors. The EGFR or Met receptor signaling positively regulates mTOR signaling thus a gain in mTOR pathway will also lead to amplification of these oncogenic signals. We noticed gain and loss of bonafide genes of mTOR pathway regulating autophagy, fatty acid metabolism and protein synthesis.

Once tumor develops to a certain size, it evolves a set of molecular machinery, which helps tumor cells to spread in the body and to identify new niche for growth. A very important molecular signaling network involved in cellular chemotaxis and strongly associated with metastasis is CXCR4-SDF1 signaling pathway. Our data set identify genes that are involved in chemotactic signaling in cancer cells transduced by CXCR4. Another important pathway associated with malignancy and metastasis is IL6 signaling pathway. Several known amplifiers and regulators annotated in this signaling pathway are significantly gained in tumor samples including STAT family members as well as upstream activators such as IL6R and JAK2. Collectively, we demonstrate that signaling pathway prediction by copy number analysis in cancer patient samples not only identifies oncogenes but also identifies the crucial oncogenic signaling pathways. The benefit of our approach lies in exhaustive analysis of cancer samples by integrating the information generated by computational analysis with intuitive networking of identified pathways for a biological output. Using this approach one can nail down putative drug targets as well biomarkers for cancer management.

Conclusion

Systematic analysis of genomic CNA can provide information about genes involved in cancer initiation and progression. While focusing on recurring CNA hot-spots has been useful in identifying candidate oncogenes, increasingly the complex contributions of multiple, cooperating genes and pathways has come into focus. With the commonly observed involvement of large genomic regions, CNA can affect the expression of a multitude of genes far beyond canonical oncogenes or tumor suppressors. This argument is substantiated by the observation of recurring patterns of cancer type specific CNA which point towards tumor-dependent distinct evolutionary pressure on large scale CNA selection.

In our view, CNA data can be used to identify relevant functional cooperations of multiple genes in a given tumor type. We base our approach on the identification of pathways whose member's genomic loci are enriched among CNA. Affected pathways allow the identification of tumor type related molecular mechanisms and, by conjecture, their probable functional relevance in those cancers. In contrast to other attempts, our analysis is purely based on CNA data, which is increasingly becoming available for large patient cohorts and does not suffer from the experimental and platform-related normalization problems inherent to e.g. expression data series. Our analysis also tests the hypothesis if and how CNA can be used to detect pathway alterations in cancer.

We here give a detail analysis of the tools which can be used to find CNA targeted signaling pathways in

cancer. According to our analysis CNA harbor several genes and can in turn affect several pathways. When the genomic clustering of pathways is considered then a gene based analysis to look for enriched pathways does not produce any significant results. The segmentation approach (S-path) identifies pathways altered in CNA data and considers the fact that pathways can be clustered on the genome. However it does not solve the issue of identifying which genes could be important for a clustered set of genes for any pathway. In our analysis of various unrelated cancer types, we were able to identify common signaling pathways that likely have pathogenetic relevance in these entities. The results of our analyses point towards a convergence of distinct signaling towards a restricts set of interconnected cellular responses. While considering a multitude of CNA-affected, potentially functionally relevant genes, the discrimination of a limited set of effector pathways specific to each entity may help to provide guidance for selecting appropriate therapeutic targets.

In our opinion, a bioinformatics methodology based on large datasets, in combination with informed biological analysis can help to define future modes of cancer management. We expect that the S-path method will prove to be a powerful utility for large scale, comparative analyses of cancer genome datasets.

Conclusions

Author's contributions

NK, HC, MB collected the data. NK and MB have designed the analysis. NK performed all the analysis. NK,HC and MB wrote and corrected the manuscript.

Acknowledgements

NK is supported through a grant by Krebsliga Schweiz (Swiss Cancer League) and University Research Priority Program (URPP) at University of Zurich. Haoyang Cai is supported through a grant from China Scholarship Council.

References

1. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VEH, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, Sauvage FJD, Faham M, Seshagiri S: **Diverse somatic mutation patterns and pathway alterations in human cancers**. *Nature* 2010, **466**(7308):869–873.
2. Baudis M: **Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data**. *BMC Cancer* 2007, **7**:226.
3. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Henry KTM, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W,

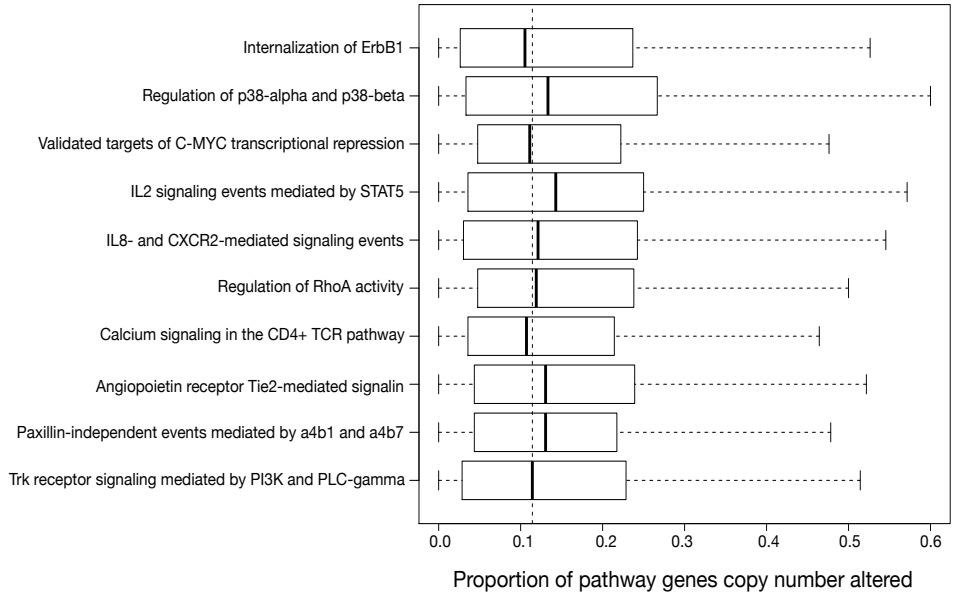


Figure 1: Proportion of genes showing CNA across entire cancer CNA data; 10 randomly chosen pathways are shown here. The vertical line represents the average number of genes being hit (median of per-case fraction of window bins overlapping CNA). The overall fraction of window bins being hit in a given pathway fluctuates around the expected value.

Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M: **The landscape of somatic copy-number alteration across human cancers**. *Nature* 2010, **463**(7283):899–905.

4. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control**. *Nat Med* 2004, **10**(8):789–799.
5. Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, Mclendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Hartigan J, Smith DR, Strausberg RL, Marie SKN, Shinjo SMO, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW: **An Integrated Genomic Analysis of Human Glioblastoma Multiforme**. *Science* 2008, **321**(5897):1807–1812.
6. Subramaniana A, Tamayoa P, Moothaa VK, Mukherjeed S, Eberta BL, Gillettea MA, Paulovichg A, Pomeroyh SL, Goluba TR, Landera ES, Mesirova JP, **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):6.
7. Edelman EJ, Guinney J, Chi JT, Febbo PG, Mukherjee S: **Modeling Cancer Progression via Pathway Dependencies**. *PLoS Comput Biol* 2008, **4**(2):e28.
8. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA, Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**. *Nature* 2006, **439**(7074):353–357.

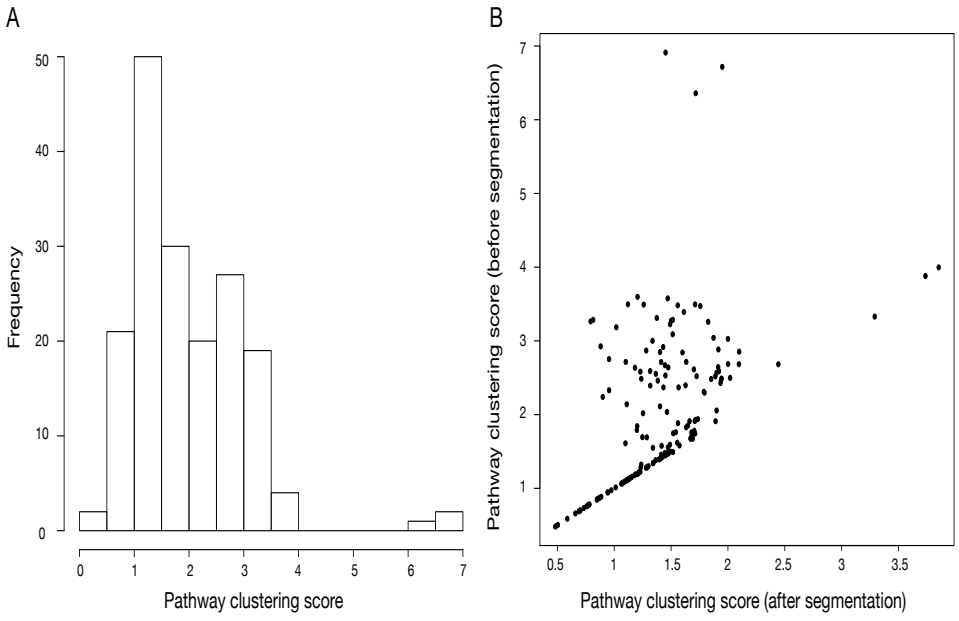


Figure 2: A) The distribution of scores obtained for the genomic clustering of pathway genes. log10 of the scores are plotted here for simplicity. B) Comparison of pathway clustering score before and after segmentation. Dividing the genome into small segments reduces the pathway clustering scores.

9. Kim TM, Yim SH, Jeong YB, Jung YC, Chung YJ: **PathCluster: a framework for gene set-based hierarchical clustering**. *Bioinformatics* 2008, **24**(17):1957–1958.

10. Shen K, Tseng GC: **Meta-analysis for pathway enrichment analysis when combining multiple genomic studies**. *Bioinformatics* 2010, **26**(10):1316–1323.

11. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A, Nevins JR: **A pathway-based classification of human breast cancer**. *Proceedings of the National Academy of Sciences* 2010, **107**(15):6994–6999.

12. Mclendon R, Friedman A, Bigner D, Meir EGV, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, Yung WKA, Bogler O, Vandenberg S, Berger M, Prados M, Muzny D, Morgan M, Scherer S, Sabo A, Nazareth L, Lewis L, Hall O, Zhu Y, Ren Y, Alvi O, Yao J, Hawes A, Jhangiani S, Fowler G, Lucas AS, Kovar C, Cree A, Dinh H, Santibanez J, Joshi V, Gonzalez-Garay ML, Miller CA, Milosavljevic A, Donehower L, Wheeler DA, Gibbs RA, Cibulskis K, Sougnez C, Fennell T, Mahan S, Wilkinson J, Ziaugra L, Onofrio R, Bloom T, Nicol R, Ardlie K, Baldwin J, Gabriel S, Lander ES, Ding L, Fulton RS, Mclellan MD, Wallis J, Larson DE, Shi X, Abbott R, Fulton L, Chen K, Koboldt DC, Wendl MC, Meyer R, Tang Y, Lin L, Osborne JR, Dunford-Shore BH, Miner TL, Delehaunty K, Markovic C, Swift G, Courtney W, Pohl C, Abbott S, Hawkins A, Leong S, Haipek C, Schmidt H, Wiechert M, Vickery T, Scott S, Dooling DJ, Chinwalla A, Weinstock GM, Mardis ER, Wilson RK, Getz G, Winckler W, Verhaak RGW, Lawrence MS, O'kelly M, Robinson J, Alexe G, Beroukhim R, Carter S, Chiang D, Gould J, Gupta S, Korn J, Mermel C, Mesirov J, Monti S, Nguyen H, Parkin M, Reich M, Stransky N, Weir BA, Garraway L, Golub T, Meyerson M, Chin L, Protopopov A, Zhang J, Perna I, Aronson S, Sathiamoorthy N, Ren G, Yao J, Wiedemeyer WR, Kim H, Kong SW, Xiao Y, Kohane IS, Seidman J, Park PJ, Kucherlapati R, Laird PW, Cope L, Herman JG, Weisenberger DJ, Pan F, Berg DVD, Neste LV, Yi JM, Schuebel KE, Baylin SB, Absher DM, Li JZ, Southwick A, Brady S, Aggarwal A, Chung T, Sherlock G, Brooks JD, Myers RM, Spellman PT, Purdom E, Jakkula LR, Lapuk AV,

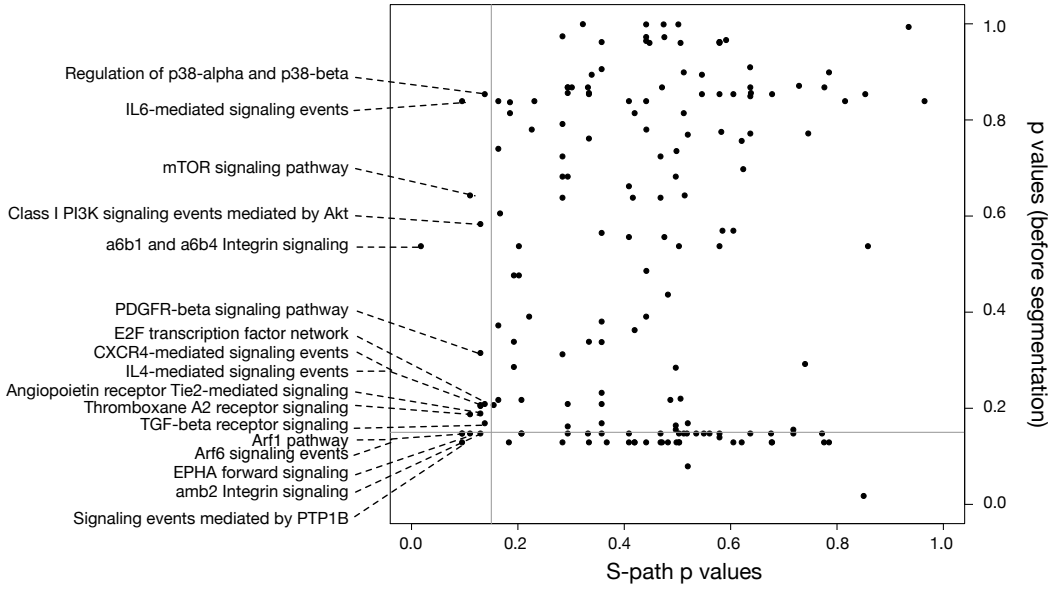


Figure 3: Bejamini-Hochberg(BH) corrected p-values of pathways enriched using S-path are plotted against BH corrected p values for clustering of those pathways. The grey lines represent the p value cutoff of 0.15 for both S-path enriched pathways and genomic clustering score. Some pathways enriched using S-path are clustered on the genome however this clustering can be ignored as none of these pathways are significantly clustering when segmentation was done.

Marr H, Dorton S, Choi YG, Han J, Ray A, Wang V, Durinck S, Robinson M, Wang NJ, Vranizan K, Peng V, Name EV, Fontenay GV, Ngai J, Conboy JG, Parvin B, Feiler HS, Speed TP, Gray JW, Brennan C, Socci ND, Olshen A, Taylor BS, Lash A, Schultz N, Reva B, Antipin Y, Stukalov A, Gross B, Cerami E, Wang WQ, Qin LX, Seshan VE, Villafania L, Cavatore M, Borsu L, Viale A, Gerald W, Sander C, Ladanyi M, Perou CM, Hayes DN, Topal MD, Hoadley KA, Qi Y, Balu S, Shi Y, Wu J, Penny R, Bittner M, Shelton T, Lenkiewicz E, Morris S, Beasley D, Sanders S, Kahn A, Sfeir R, Chen J, Nassau D, Feng L, Hickey E, Zhang J, Weinstein JN, Barker A, Gerhard DS, Vockley J, Compton C, Vaught J, Fielding P, Ferguson ML, Schaefer C, Madhavan S, Buetow KH, Collins F, Good P, Guyer M, Ozenberger B, Peterson J, Thomson E: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061–1068.

13. Lucas JE, Kung HN, Chi JTA, Tucker-Kellogg G: **Latent Factor Analysis to Discover Pathway-Associated Putative Segmental Aneuploidies in Human Cancers**. *PLoS Comput Biol* 2010, **6**(9):e1000920.

14. Holden M, Deng S, Wojnowski L, Kulle B: **GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies**. *Bioinformatics* 2008, **24**(23):2784–2785.

15. O'dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A: **The SNP ratio test: pathway analysis of genome-wide association datasets**. *Bioinformatics* 2009, **25**(20):2762–2763.

16. Lee JM: **Genomic Gene Clustering Analysis of Pathways in Eukaryotes**. *Genome Research* 2003, **13**(5):875–882.

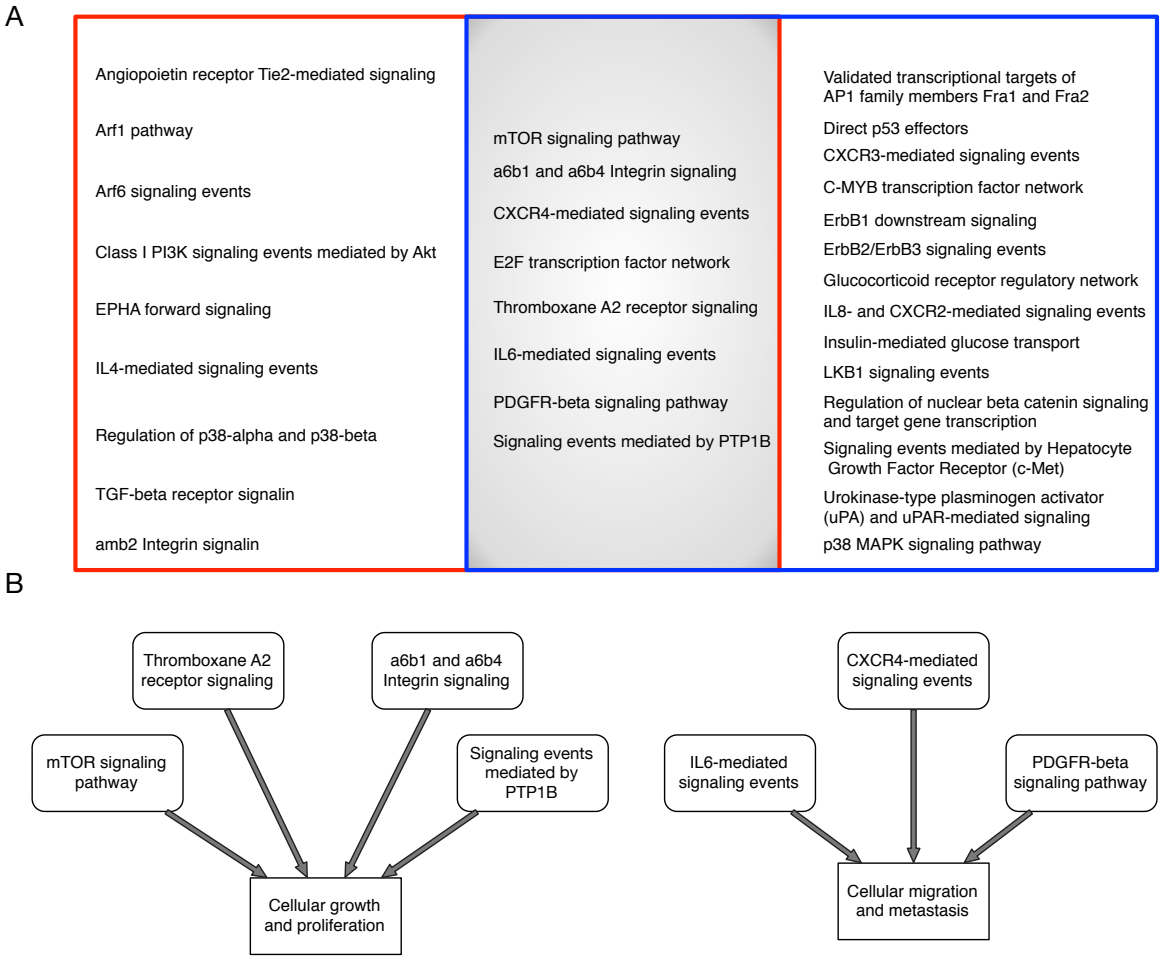


Figure 4: A) Pathways enriched by a combinatorial analysis are shown in the red rectangle and the ones enriched in multiple cancer types (20) are shown in the blue rectangle. In the intersection of both rectangles (grey shaded region) are pathways which were looked at in details. B) Network of pathways enriched in cancers. different enriched pathways (shown in rectangle with round corners) converge to affect cancer-associated cellular responses (rectangle with smooth corners).

17. Baudis M, Cleary ML: **Progenetix.net: an online repository for molecular cytogenetic aberration data**. *Bioinformatics* 2001, **17**(12):1228–9.

18. Joshi-Tope G: **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Research* 2004, **33**(Database issue):D428–D432.

19. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Research* 2010, **38**(Database):D355–D360.

20. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic Acids Research* 2009, **37**(Database):D674–D679.

21. Curtis RK, Orešič M, Vidal-Puig A: **Pathways to the analysis of microarray data**. *Trends in Biotechnology* 2005, **23**(8):429–435.

22. Guo W, Pylayeva Y, Pepe A, Yoshioka T, Muller WJ, Inghirami G, Giancotti FG: **4 Integrin Amplifies ErbB2 Signaling to Promote Mammary Tumorigenesis**. *Cell* 2006, **126**(3):489–502.

23. Trusolino L, Bertotti A, Comoglio PM: **A signaling adapter function for alpha6beta4 integrin in the control of HGF-dependent invasive growth**. *Cell* 2001, **107**(5):643–54.

24. Dubé N, Cheng A, Tremblay ML: **The role of protein tyrosine phosphatase 1B in Ras signaling**. *Proc Natl Acad Sci USA* 2004, **101**(7):1834–9.

25. Klingler-Hoffmann M: **The Protein Tyrosine Phosphatase TCPTP Suppresses the Tumorigenicity of Glioblastoma Cells Expressing a Mutant Epidermal Growth Factor Receptor**. *Journal of Biological Chemistry* 2001, **276**(49):46313–46318.

26. Mattila E, Pellinen T, Nevo J, Vuoriluoto K, Arjonen A, Ivaska J: **Negative regulation of EGFR signalling through integrin-11-mediated activation of protein tyrosine phosphatase TCPTP**. *Nature* 2005, **7**:78–85.

9.3 Publication 3: Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated cytogenetic data (submitted)

Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated cytogenetic data

Nitin Kumar[1#], Haoyang Cai[1#], Christian von Mering[1,2,*] Michael Baudis[1,*]
1 Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
2 Swiss Institute of Bioinformatics, Quartier Sorge, Lausanne, Switzerland
∗ **E-mail: mbaudis@imls.uzh.ch, mering@imls.uzh.ch**
# **equal contribution**

Abstract

Background

Regional genomic copy number alterations (CNA) are observed in the vast majority of cancers. Besides specifically targeting well-known, canonical oncogenes, CNAs may also play more subtle roles in terms of modulating genetic potential and broad gene expression patterns of developing tumors. Any significant differences in the overall CNA patterns between different cancer types may thus point towards specific biological mechanisms acting in those cancers. In addition, differences among CNA profiles may prove valuable for cancer classifications beyond existing annotation systems.

Principal Findings

We have analyzed molecular-cytogenetic data from 25579 tumors samples, which were classified into 160 cancer types according to the International Classification of Disease (ICD) coding system. When correcting for differences in the overall CNA frequencies between cancer types, related cancers were often found to cluster together according to similarities in their CNA profiles. Based on a randomization approach, distance measures from the cluster dendrograms were used to identify those specific genomic regions that contributed significantly to this signal. This approach identified 43 non-neutral genomic regions whose propensity for the occurrence of copy number alterations varied with the type of cancer at hand. Only a subset of these identified loci overlapped with previously implied, highly recurrent (hot-spot) cytogenetic imbalance regions.

Conclusions

Thus, for many genomic regions, a simple null-hypothesis of independence between cancer type and relative copy number alteration frequency can be rejected. Since a subset of these regions display relatively low overall CNA frequencies, they may point towards second-tier genomic targets that are adaptively relevant but not necessarily essential for cancer development.

Introduction

Genetic changes such as point mutations, regional copy number alterations/aberrations (CNA) and structural changes (e.g. gene fusion events) are all hallmarks of cancer. CNAs arise as somatic changes in the tumor cell genome through a variety of mechanisms and can be observed in virtually all types of cancer, to a varying extent. So far, the most widely used methods for the detection of CNAs have been chromosomal and array-based Comparative Genomic Hybridization (CGH) techniques [1–4]. Localized, recurring CNAs (hot-spots) have been shown to target canonical oncogenes (e.g. duplications/amplifications of the MYC, MYCN, REL loci) or tumor suppressor genes (e.g. deletions of the CDKN2A/B, TP53, ATM

loci). Some regional CNAs such as gains on 8q and losses on 3p are present across multiple cancer types, whereas other imbalances may be largely restricted to a limited number of cancer entities [5].

Datasets integrated across multiple cancer types have previously been analyzed, to report regional "hot-spots" of frequent CNAs [5, 6]. In a given set of individual tumor samples, the number and distribution of CNAs varies considerably [5] and this genetic heterogeneity has been used to detect and report co-occurring CNAs [7].

In principle, specific patterns and similarities in the individual and/or disease specific CNA profiles might point to distinct oncogenomic mechanisms acting in different cancer types and specimens, given a sufficiently large number of data points. Indeed, clustering of CNA patterns has been used to identify oncogenomic similarities [5,8–11]. The adaptation of clustering techniques to the analysis of CNA patterns has been subject of previous studies [12–14]. With a few exceptions [5, 14], however, sample-based clustering has been the main focus of such studies so far. In contrast, we here explore the clustering of cancer types, not of individual cancer samples.

Both descriptive and clustering-based analyses of CNA across multiple cancer types suffer from a bias towards the more frequently occurring events. Due to the heterogeneity of the overall CNA signal, with greatly varying average frequencies of CNAs per cancer type (Figure 1a), clustering results may be distorted depending on the disease entities analyzed. This variation in overall CNA occurrence frequencies across cancer types may simply be owed to differences in the average time points of clinical detection or in different progression characteristics, and should be corrected for prior to clustering analyses. To the best of our knowledge, so far no implementation has been reported for a comprehensive, very large-scale clustering analysis of frequency-normalized cancer CNA profiles.

Here, we focus on the identification of genomic regions that contribute meaningfully to the clustering of cancer types. From hereon we will refer to those as "non-neutral" regions. As the starting point of our analysis, we use hierarchical clustering to arrange cancer types on the basis of their CNA frequency profiles. We then employ a permutation approach to estimate the relative contribution of individual genomic regions to the quality of the clustering and to the derived relationship tree. The clustering quality is inferred from an intrinsic measure (summed branch lengths: tree height statistics), and genomic regions that reject the null hypothesis are termed non-neutral. Identified regions are compared to canonical CNA hot-spots (i.e. those that occur most frequently across the entire dataset).

Our current analysis is based on data from a total of 25579 samples, which are classified into 160 different cancer entities (supplement table) according to the International Classification of Disease in Oncology (ICD-O 3). Our approach is unique in that it a) focuses less on the clustering as such but more on the individual genomic regions that best support the clustering, b) uses an intrinsic quality measure coupled to a permutation strategy for validation, c) performs CNA frequency normalization prior to analysis, and d) is based on a very large data set, processed in a standardized setup. We aim for the identification of potential cancer-specific driver/modulator regions, which may not have been detected in earlier, largely hot-spot-focused approaches. All of the underlying cancer data is available through our Progenetix repository (www.progenetix.org; [15]).

Results

The average overall frequency of CNAs across the entire genome varies among different cancer types (Figure 1a). Since the relative weight of CNAs at individual genomic regions in a given cancer type depends on the observed overall genome-wide frequency, we aggregated all patient samples by cancer type and normalized the frequencies of CNAs for each cancer type to the overall mean observed across the entire data set (Figure 1b, Figure S1). The normalized CNA frequency profiles were then clustered using hierarchical clustering. To evaluate the quality and the biological signal in the clustering, we labeled each cancer type with its "root" cell type (i.e., an undifferentiated cell type from which the tumor likely originated). We expected cancers of the same root cell type to cluster together; this was

used as an external proxy for the expected biological relationships between cancer entities. The Random Index [16] was used to compute this external cluster quality measure. Tumors of the same cell type indeed often clustered together, usually in 2-3 small groups (Figure 2). The consistency of this grouping was significantly higher than expected at random, pointing towards biologically meaningful differences in CNA profiles between tumors of distinct origins. Cutting the tree at several heights always led to an observed quality of clustering that was better than the expected random value (Figure 2), except for the cut at the highest level, which resulted in only three clusters. This strongly argues against a completely neutral occurrence pattern of CNAs in the genome, and supports a correlation between biologically meaningful groups of cancer entities and their CNA profiles. Randomizations of the entire frequency matrix lead to a complete loss of the signal present in the clustering tree (Figure S2), and also strongly reduced the summed branch lengths tree-height statistic.

Non-neutral CNAs

The normalized and clustered frequency matrix encompassing 160 large-scale genomic regions and 160 cancer types is shown in Figure 3. To determine how much each individual genomic region contributes to the overall signal, we individually randomized its profile across cancer types, while keeping the rest of the data unchanged. We then examined the concomitant reduction in the tree length statistics (TLS) of the clustering dendrogram, upon 100000 independent randomizations, to determine the statistical significance of that region's contribution. The resulting cancer-diverging CNA regions are important as they cannot be fully neutral and have the potential to define relationships among cancer types. Indeed, 43 out of the 160 genomic regions (supplemental table) were observed to have a non-neutral contribution (Bonferroni-corrected p-value ≤ 0.016) in the aggregated cancer CNA data. Note that gain and loss events were treated independently, and no preferential bias towards gains or losses was observed among the detected non-neutral regions (22 gains and 21 losses). The CNA occurrence frequencies of the non-neutral genomic regions spread thorough the entire frequency spectrum (Figure 4). Only 13 (8 gains and 5 losses) of the non-neutral regions were found altered overall more often than average (Figure 5, intersection of black and grey rectangle), indicating that subset of frequently altered hotspot regions carry a detectable signal to distinguish cancer types (the number of frequently altered regions stands at 59; Bonferroni-corrected p-value ≤ 0.016, supplementary table). This observation emphasizes our key point that not only the frequent CNA regions should be used to cluster and annotate cancer types.

22 genomic intervals across 12 chromosomes were found to be informative when specifically considering duplications/gains only (Table 1 and Figure 5). All three genomic segments of chromosome 18 (18p1, 18p2, 18q2) exhibited a signal. For other chromosomes such as chromosome 1 (1q2,1q3,1q4,1p2), chromosome 3 (3q1, 3q2, 3p1), chromosome 12 (12q1,12q2) and chromosome 21 (21p1, 21q1) more than 50% of genomic regions were informative as gains, suggesting simultaneous involvement of multiple loci from these chromosomes. Changes on chromosome 1(1p2), chromosome 3 (3p1, 3q1), chromosome 5 (5q2, 5q3), chromosome 9 (9p1), chromosome 11 (11p1), chromosome 12 (12q1, 12q2), chromosome 18 (18p1, 18q1, 18q2) and chromosome 21 (21p1, 21q1) were selectively informative only as gains. In terms of deletions/losses, 10 chromosomes encompassing 21 genomic regions were found to be non-neutral. Like for chromosome 18 gains, the complete chromosome 7 (7p1, 7p2, 7q1, 7q2, 7q3) was found to be informative when lost (Table 1). Informative regions on chromosome 1 (1p1,1q1, 1q2, 1q3, 1q4) and chromosome 9 (9q1, 9q3, 9p2) covered more than 50% of genomic segments present on these chromosomes. Selective losses were observed on chromosome 1 (1p1, 1q1), chromosome 6 (6q2), 7 (7q1, 7q2, 7q3, 7p2) , 8 (8q1, 8q2), 9 (9p2, 9q1, 9q3), 12 (12p1), 16 (16q1). CNAs involving chromosome 1 (1q2, 1q3, 1q4), chromosome 3 (3q2), chromosome 7 (7p1), chromosome 19 (19p1) and chromosome 22 (22q1) were informative both as gain and loss events. This represents a small proportion (16%) of non-neutral CNA. Involvement of a region both as gain and loss may point towards multiple adaptively relevant loci, and/or towards a generally unstable nature of these regions.

Cancer diverging nature of non-neutral CNA

To provide few examples of cancer classifying behavior of non-neutral changes, we selected a few of the enriched changes and analyzed them for their specific occurrence in different cancers. An example include cancer entities showing predominant losses versus gains on 7q. Preferential losses involving 7q were observed in germ cell, myeloid and myeloproliferative tumors (Figure 3) whereas neuroepithelial brain tumors (among other entities) preferentially displayed gains on 7q. Losses involving 7q are common in myeloid and myeloproliferative tumors [17–20] and are associated with advanced age and resistance to therapies [21, 22]. However here we show that 7q losses are quite specific to myeloid tumors and give them a selective divergence from rest of the cancer types. 7q losses in germ cell tumors are not being looked at in details [23, 24]. Our analysis provides a strong link between myeloid and germ cell tumors, linking them with respect to their 7q profiles and differentiating them from other cancer types such as neuroepithelial brain tumors showing a preferential gains of 7q.

Chromosome 8q gains are very common in most of the cancers [5, 6]. However 8q losses were enriched as non-neutral in our analysis. Preferential losses involving 8q were present in some brain tumors (medulloblastoma, Figure 3), diverging them from other epithelial tumors. Differences in preferential losses involving 8q separated neuroepithelial tumors in two categories both having gains on 7q but only one (mainly meduloblastomas) having preferential losses on 8q (Figure S3). Losses involving chromosome 8q across medulloblastomas have been reported by a few [25] studies but have not been looked at in details. Our analysis shows that 8q losses are selective for medulloblastomas and can be important for cancer development/progression. Preferential losses of 8q were also observed in germ cell tumors separating them from epithelial tumors (Figure S4).

To give another examples we also looked for cancers showing gains involving chromosome 18. Follicular lymphomas exhibited specific gains on chromosome 18 where as epithelial tumors prefered to loose chromosome 18 (Figure S4). Chromosome 18 gains are very common in follicular lymphomas [26, 27] however their role in separating them from other cancer types is shown here.

Discussion

Our current study represents the largest analysis performed to date on cancer CNA data, with the aim of detecting oncogenomic features that may be specifically associated or enriched in certain subsets of cancer entities. In contrast to gene-centric approaches, our analysis assesses the complete information space of genomic copy number imbalances from whole genome profiling experiments.

Overall, the frequency of CNAs across genomic intervals varied between between 0.01% to 23% (Figure 4). Clustering of cancer types on the basis of their frequency profiles helped to identify a class of underlying molecular signals that is orthogonal to histological classifications or clinical categories (the latter are predominantly driven by the affected organ/tissue). Cancer types vary from each other in their CNA abundance, CNA size spectrum and degree of genomic instability. With respect to genomic coverage, large CNAs are generally frequent in cancer [6] and should not be excluded from statistical analyses of cancer genome patterns. While comparing CNA profiles of cancer types, their complexity and variation in frequencies have to be considered. When correcting for these parameters, regional CNAs defining the divergence of the overall profiles can be delineated.

We performed an analysis of a global cancer CNA dataset, identifying 43 genomic regions on 15 chromosomes as significant for CNA profile divergence in cancer types. Obviously, these changes do not cover the entire spectrum of CNA events in cancer, but define a subset of genomic regions that may have a possibly adaptive link to the distinct biology of various cancer types. These regions overlap rather poorly with hotspot regions observed in many cancers. This suggests that hot-spot regions, though frequently associated with canonical oncogenes, may not always be very useful in aiding data-driven evaluation of cancer (sub-)types.

With our current study, we aim to provide a new, generalized approach at identifying genomic elements relevant in the genesis of individual cancer entities. Though here showcasing a "global" approach without entity pre-selection, our methodology may prove valuable when targeting relevant genomic separators in limited, biologically related entity sets. Since the current analysis is based primarily on molecular-cytogenetic data from chromosomal CGH experiments with a spatial resolution of several megabases, only inferred information about the causal genes present in the non-neutral regions could be obtained. With upcoming high-resolution genomic array and/or sequencing data, similar analyses will more specifically define the non-neutral CNAs and can be valuable starting points for an integration of the results with functional pathway frameworks. Also, although we have focused our current analysis solely on a CNA dataset, our approach should prove particularly valuable when combined with other sets of related diagnostics (for example point mutation data), whereby the assignment of possible driver genes in the non-neutral regions might become feasible.

Materials and Methods

Data

Cancer CNA data from chromosomal and array CGH experiments across 160 cancer types (a total of 25579 samples), as classified on the basis of International Classification of Disease codes (ICD), have been collected over the last decade [5]. For our analysis, regional CNA information across all cancer types was reduced to 80 genomic intervals covering the entire genome with the exception of the sex chromosomes. Gain and loss events were considered separately for the analysis, resulting in a matrix of dimensions nXm, where n is the number of samples and m is the number of genomic intervals (*i.e.* 160).

Cancer clustering

The frequency of CNA changes across all genomic intervals was computed for each ICD type, and the entire frequency matrix was then normalized (Figure S1). The frequency matrix was ordered using hierarchical Ward clustering. The aggregated separation distance between cancer entities obtained using hierarchical clustering can be analyzed by parsing the clustering tree (dendrogram). The tree represents the relatedness among groups present in the same clade (similar to phylogenetic trees). Randomized data disrupts the tree completely (Figure S2), and the overall tree height statistic is reduced 3-fold, reflecting the complete loss of ordering information present in the original tree.

Method to compare tree height

We used the tree height as an intrinsic measure to compare cancer associations obtained using clustering and to gauge the information present in the tree; this was used to define non-neutral CNAs. This has advantages over traditional clustering evaluation techniques, as it a) does not require external gold standard information, and b) does not require cutting the tree at an arbitrary distance. The overall tree height is defined as the sum of all direct parent-child relation path lengths in the tree. Tree distances (branch lengths) generally reflect the CNA profile discrepancies between two cancers (or groups of cancers). For any node i, the tree height between this node and its immediate parent j can be measured as $TH_j - TH_i$. The overall tree height of a tree with n nodes is than obtained as $OTH = \sum_{i=1,j=1}^{i=n,j=n} TH_j - TH_i$ (supplement figure 3).

Tree length statistics (TLS)

To identify genomic regions that are non-neutrally affected by CNA we have developed the following permutation strategy:

1. Normalized frequencies of CNA across all genomic intervals are computed across all cancer types.
2. The cancer classification tree is obtained using hierarchical Ward clustering.
3. The observed over all tree height (OTH_o) is calculated as mentioned above (Figure S5).
4. A counter C is set to zero for every genomic interval in consideration.
5. For any genomic interval i, its status values are shuffled among all samples keeping its over all frequency the same (n_i).
6. The frequency of CNA at genomic interval i is re-calculated after randomization across all cancer types. The shuffling in the previous step changes the frequency of interval i across all cancer types keeping the normalized frequency distribution of all other genomic intervals.
7. The frequencies for interval i in the normalized frequency matrix from step one are replaced with permuted frequencies for this interval and the permuted overall tree heigh ($OTH_{i,p}$) is computed.
8. If $OTH_{i,p} \geq OTH_o$, C is incremented as C = C + 1.
9. p-value for genomic location i, at the end of N (100'000) permutations is computed as $p_i = C/N$.
10. p-values across all bands are corrected for false discovery rate using Bonferroni correction.

Frequency based enrichment (FBE)

Frequently observed CNA regions ("hot-spots") are genomic changes that occur more often than expected under a fully random null model. Such hot-spot CNAs can be identified using the binomial probability function [28]. Let's suppose genomic interval i shows a CNA across n_i samples out of N samples. The background CNA frequency (n_b) can be represented as the mean frequency change across all intervals. The p value that the frequency of CNA n_i, is more than any frequency x ($n_i \geq x$) is obtained using the binomial probability function.

$$p(n_i|N, n_b) = \binom{N}{n_i} n_b^{n_i} (1 - n_b)^{N - n_i}$$

$$p_i = \sum_{n=x}^{N} p(n_i|N, n_b)$$

Genomic intervals showing a large deviation from the mean will be assigned low p-values. All p-values are corrected for false discovery rate using Bonferroni correction

Acknowledgments

References

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258: 818–21.
2. Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, et al. (1993) Detection of amplified dna sequences by reverse chromosome painting using genomic tumor dna as probe. Hum Genet 90: 584–9.
3. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, et al. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer 20: 399–407.

4. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–11.

5. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal cgh data. BMC Cancer 7: 226.

6. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

7. Kumar N, Rehrauer H, Cai H, Baudis M (2011) Cdcoca: a statistical method to define complexity dependence of co-occuring chromosomal aberrations. BMC Med Genomics 4: 21.

8. Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, et al. (2006) Dna copy number amplification profiling of human neoplasms. Oncogene 25: 7324–7332.

9. Liu J, Ranka S, Kahveci T (2007) Markers improve clustering of cgh data. Bioinformatics 23: 450–7.

10. Ferreira BI, Garcia JF, Suela J, Mollejo M, Camacho FI, et al. (2008) Comparative genome profiling across subtypes of low-grade b-cell lymphoma identifies type-specific and common aberrations that target genes with a role in b-cell neoplasia. Haematologica 93: 670–679.

11. Takeuchi I, Tagawa H, Tsujikawa A, Nakagawa M, Katayama-Suguro M, et al. (2009) The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of b-cell lymphomas and genetic regions involved in lymphomagenesis. Haematologica 94: 61–69.

12. Liu J, Ranka S, Kahveci T (2006) Markers improve clustering of cgh data. Bioinformatics 23: 450–457.

13. Wieringen WNV, Wiel MAVD, Ylstra B (2008) Weighted clustering of called array cgh data. Biostatistics 9: 484–500.

14. Liu J, Bandyopadhyay N, Ranka S, Baudis M, Kahveci T (2009) Inferring progression models for cgh data. Bioinformatics 25: 2208–15.

15. Baudis M, Cleary ML (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics 17: 1228–9.

16. Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. Bosotn, MA, USA: Addison Wesley.

17. Kühn MWM, Radtke I, Bullinger L, Goorha S, Cheng J, et al. (2012) High-resolution genomic profiling of adult and pediatric core-binding-factor acute myeloid leukemia reveals new recurrent genomic alterations. Blood .

18. Woo KS, Kim KE, Kim KH, Kim SH, Park JI, et al. (2009) Deletions of chromosome arms 7p and 7q in adult acute myeloid leukemia: a marker chromosome confirmed by array comparative genomic hybridization. Cancer Genet Cytogenet 194: 71–4.

19. Cordoba I, González-Porras JR, Nomdedeu B, Luño E, de Paz R, et al. (2012) Better prognosis for patients with del(7q) than for patients with monosomy 7 in myelodysplastic syndrome. Cancer 118: 127–133.

20. Aktas D, Tuncbilek E (2006) Myelodysplastic syndrome associated with monosomy 7 in childhood: a retrospective study. Cancer Genet Cytogenet 171: 72–5.

21. Appelbaum FR, Gundacker H, Head DR, Slovak ML, Willman CL, et al. (2006) Age and acute myeloid leukemia. Blood 107: 3481–5.

22. Wong JCY, Zhang Y, Lieuw KH, Tran MT, Forgo E, et al. (2010) Use of chromosome engineering to model a segmental deletion of chromosome band 7q22 found in myeloid malignancies. Blood 115: 4524–32.

23. McIntyre A, Summersgill B, Lu YJ, Missiaglia E, Kitazawa S, et al. (2007) Genomic copy number and expression patterns in testicular germ cell tumours. Br J Cancer 97: 1707–12.

24. Veltman I, Veltman J, Janssen I, van de Kaa CH, Oosterhuis W, et al. (2005) Identification of recurrent chromosomal aberrations in germ cell tumors of neonates and infants using genomewide array-based comparative genomic hybridization. Genes Chromosomes Cancer 43: 367–76.

25. jing Sun Y, zhu Yu S, yun Sun C, Wang Q, mei Jin S, et al. (2010) [detection of chromosomal dna imbalance in medulloblastoma by comparative genomic hybridization]. Zhonghua Bing Li Xue Za Zhi 39: 606–10.

26. Cheung KJJ, Delaney A, Ben-Neriah S, Schein J, Lee T, et al. (2010) High resolution analysis of follicular lymphoma genomes reveals somatic recurrent sites of copy-neutral loss of heterozygosity and copy number alterations that target single genes. Genes Chromosomes Cancer 49: 669–81.

27. Schwaenen C, Viardot A, Berger H, Barth TFE, Bentink S, et al. (2009) Microarray-based genomic profiling reveals novel genomic aberrations in follicular lymphoma which associate with patient survival and gene expression status. Genes Chromosomes Cancer 48: 39–54.

28. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466: 869–873.

Figure Legends

Figure 1: The overall frequency of genomic copy number alterations (CNA) differs among cancer types.

Boxplots show the CNA frequency distributions among tumor samples in 10 randomly selected cancer types. The boxplot delineations mark the percentiles 5%, 25%, 75% and 95%. The red lines indicate the mean frequency for each cancer type, whereas the blue line represents the overall mean frequency across all 160 cancer types analyzed here. Frequency values are defined as the ratio of number of samples showing a CNA for a genomic region (i.e., cytogenetic bands) over total number of samples in that cancer type. a) Before normalization b) After normalization. In b) the nominal frequency distribution for each cancer type is re-scaled so that its mean matches the overall mean across all cancer types. (NOS - "not otherwise specified": high-level classifications, not further assigned to lower, more detailed levels).

Figure 2: The tissue type of a cancer has a strong influence on its CNA likelihood pattern.

a) examples of individual chromosome segments, showing their observed CNA frequencies stratified by cell type. Each dot summarizes all samples classified under one particular ICD type, color-coded by root cell type. In the left panel, three chromosome segments are shown that exhibit strong differences between cell types; on the right, three negative examples without such a signal. All p-values were corrected for multiple testing according to Benjamini-Hochberg. b) the dendrogram (tree) has been obtained using hierarchical Ward clustering on the global frequency-normalized CNA profiles across all 160 genomic regions. Cancer types are again color-coded according to the cell type of origin, with the same legend as in a). Partitioning the tree by cutting at different heights produces multiple clusters; validation of those clusters based on the cancer origin (metric: Random Index) shows that the clustering works significantly better than expected at random.

Figure 3: Examples for non-neutral CNA regions

a) Heatmap of CNA profiles on genomic regions (same clustering as in Figure 2). Genomic locations are represented with orange color when considering duplications/gains, and in blue when considering deletions/losses. Color intensity shows relative CNA frequencies; the most-affected region in each row is arbitrarily set the to brightest color (1.0) for display purposes. b) Small regions (black rectangles on the heatmap) are zoomed in to show how non-neutral CNAs can differentiate between cancer types. The example shows that 7q is preferentially gained in brain tumors (red labels) whereas it is preferentially lost in germ cell (black labels), myeloid and myeloproliferative cancer types (blue labels). c) Small regions (red rectangles on the heatmap) are zoomed in to show how 8q is preferentially lost in medullublastomas (green labels) and is preferentially gained in epithelial tumors (pink labels). Some chromosomes consist entirely of non-neutral regions (such as chromosomes 18 and 7). Note that the spatial resolution of the CNA data on the chromosome is limited (roughly corresponding to cytogenetic band resolution).

Figure 4: Not only CNA "hotspots" are informative in cancer classification.

Genomic regions (bands) are sorted according to their overall frequency of CNAs observed. Those regions that are informative with respect to cancer type clustering are marked with arrows. a) Considering duplications (gains) b) Considering deletions (losses).

Figure 5: Comparison of non-neutral vs. hot-spot CNA.

Genomic regions affected by CNAs, either more frequently than average (black rectangle), or non-neutrally with respect to cancer-type classifications (grey rectangle). The intersection defines regions that are affected both frequently and non-neutrally. Changes are color-coded (gains in orange and losses in blue).

Tables

Table 1. Number of non-neutral regions per chromosome

Chromosome No.	No. genomic locations	Non-neutral gains	Non-neutral losses
1	7	4	5
2	5	-	-
3	4	3	1
4	4	-	-
5	4	2	-
6	4	-	1
7	5	1	5
8	4	-	2
9	5	1	3
10	3	-	-
11	3	1	-
12	3	2	1
13	4	1	-
14	4	-	-
15	3	-	-
16	3	-	1
17	3	-	-
18	3	3	-
19	2	1	-
20	2	-	-
21	3	2	-
22	2	1	-

Some chromosomes consist entirely of non-neutral regions (such as choromosomes 18 and 7). Note that the spatial resolution of the CNA data on the chromosome is limited (it roughly corresponds to the cytogenetic banding patterns).

Additional Files

Supplement figure 1: Method for CNA frequency normalization across cancer types.

All the frequencies among cancer types were normalized to the mean frequency of CAN changes across across the 160 cancer types. This normalization was achieved by multiplying the cancer-type-specific frequencies with an index A_n, whose value was calculated as shown.

Supplement figure 2: Dendrogram of a permuted frequency matrix

For this clustering, the frequencies among cancer types were permuted and then normalized. Hierarchial Ward clustering was then performed and the dendrogram tree shown was obtained. The tree height is severely affected by the permutation. In this randomized clustering, similar cancer types no longer clustered together.

Supplement figure 3: Small regions from heatmap in main Figure 3 are shown here

These regions represent gains and losses on 7q and 8q. 8q changes differentiate between two categories of brain tumors, with a subset showing preferential losses on 8q (green labels) and other rarely showing involvement of 8q locus (red label). Thus depending on 8q involvement neuroepithelial tumors can be divided in to two different categories. Both of them show 7q gains.

Supplement figure 4: Examples for non-neutral CNA regions

a) Heatmap of CNA profiles on genomic regions (same as in Figure 3). b) Small regions (red rectangles on the heatmap) are zoomed in to show how 8q is preferentially lost in in germ cell (black labels) tumors and is preferentially gained in epithelial cancer types (pink labels). c) Small regions (black rectangles on the heatmap) are zoomed in to show how 18q is preferentially gained in medullublastomas (brown labels) and is preferentially lost in epithelial tumors (pink labels). The examples here show that how two different non-neutral changes differential epithelial tumors from germ cell tumors and follicular lymphomas.

Supplement figure 5: Calculation of over all tree height.

Schematic representation of the summed branch-length tree height statistic. Overall tree height is computed by summing up the distance between all parents and child nodes. Note that the branch lengths of terminal branches ("leafs") are not considered. Overall tree height = $H_{A-C} + H_{B_D} + H_{AB} + H_E$.

SupplementTable.ods: Table with information about cancer types used in the analysis, non-neutral and hot-spot p values

The table giving details about all cancer types used in this analysis with the corresponding numbers of samples in them and the root cell type of each cancer. The table also has information about the non-neutral and hot-spot p-values obtained for all genomic bands in analysis.

Figure1
Click here to download Figure: 1.pdf

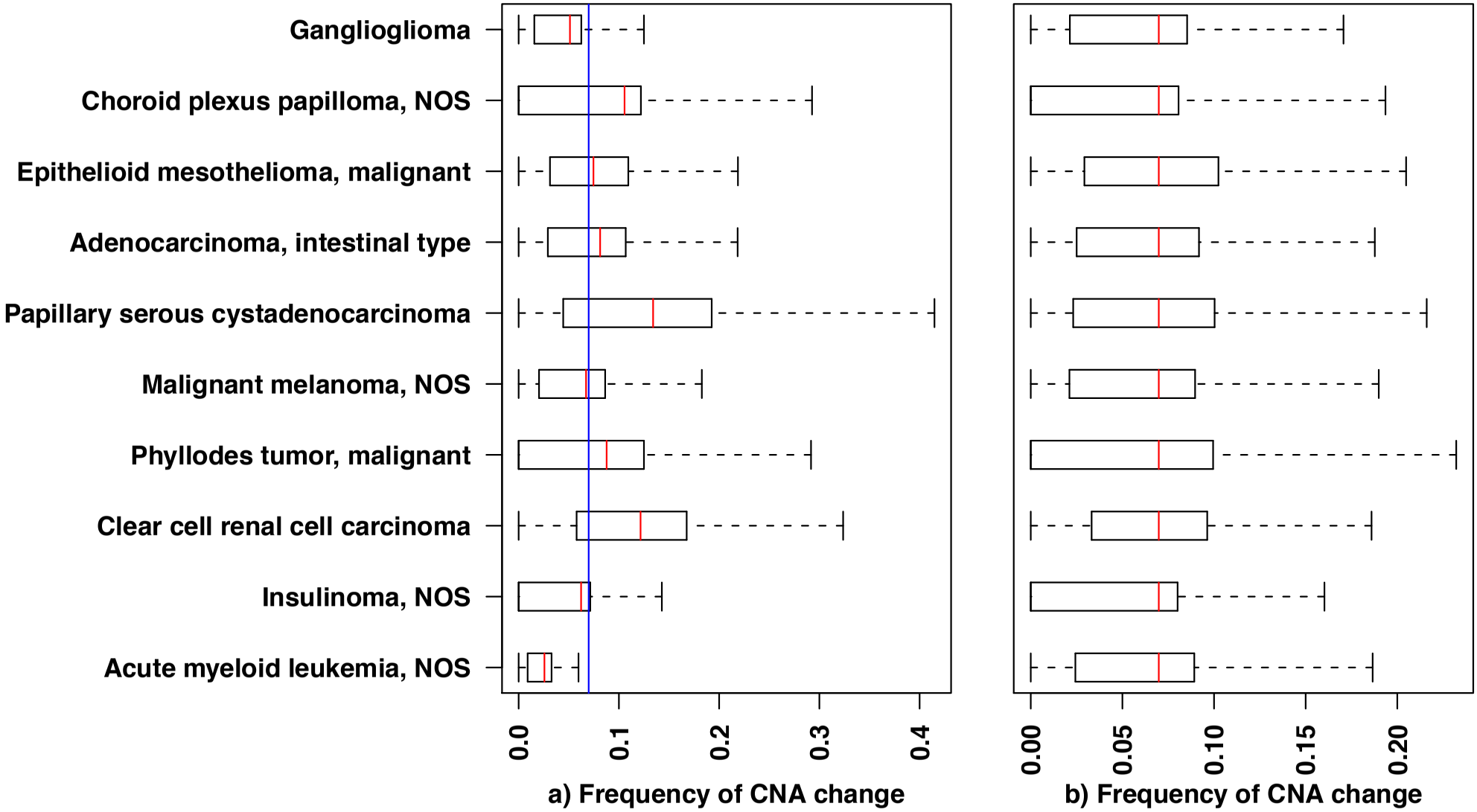


Figure 2

Click here to download Figure: 2.pdf

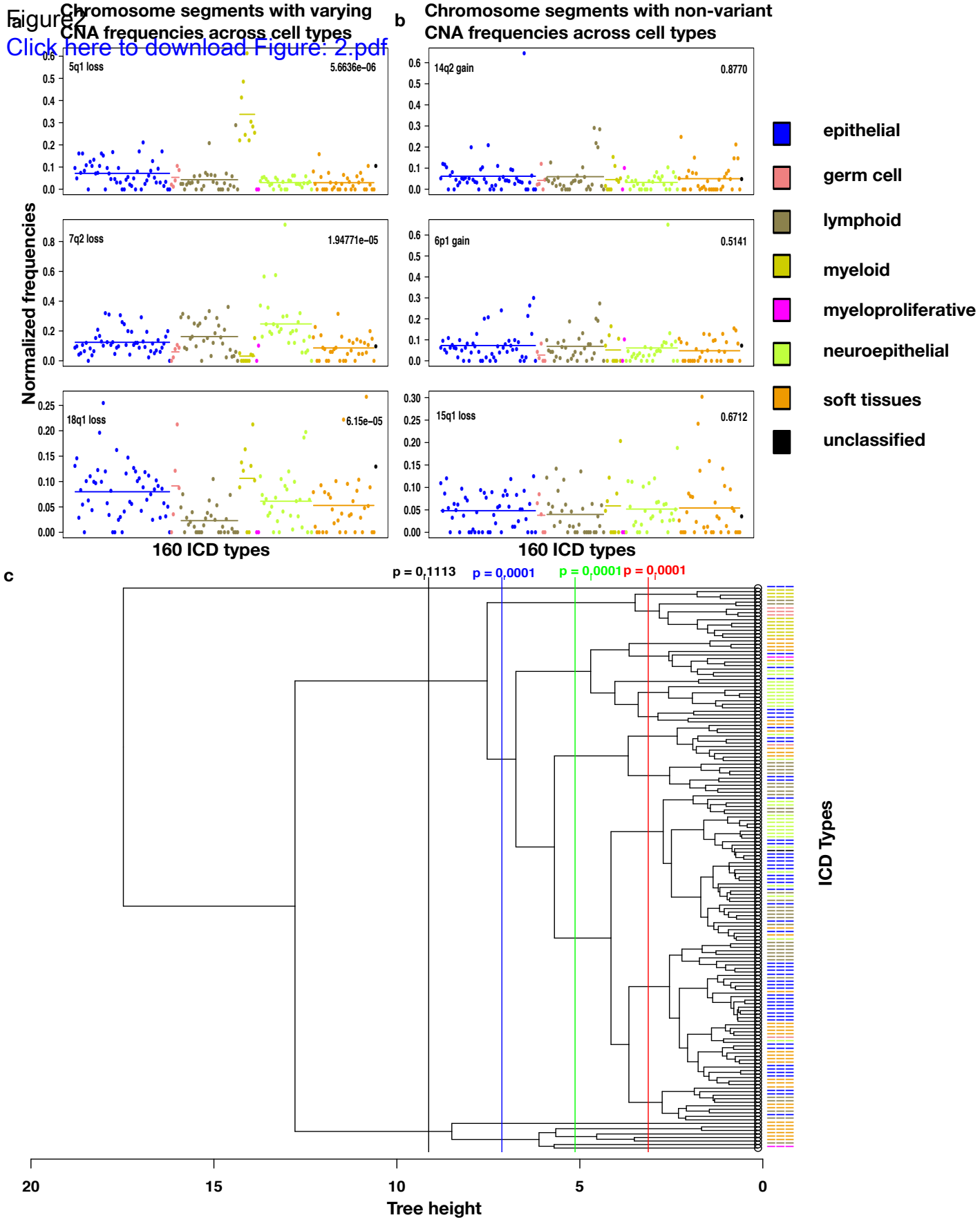


Figure3

Click here to download Figure3.pdf

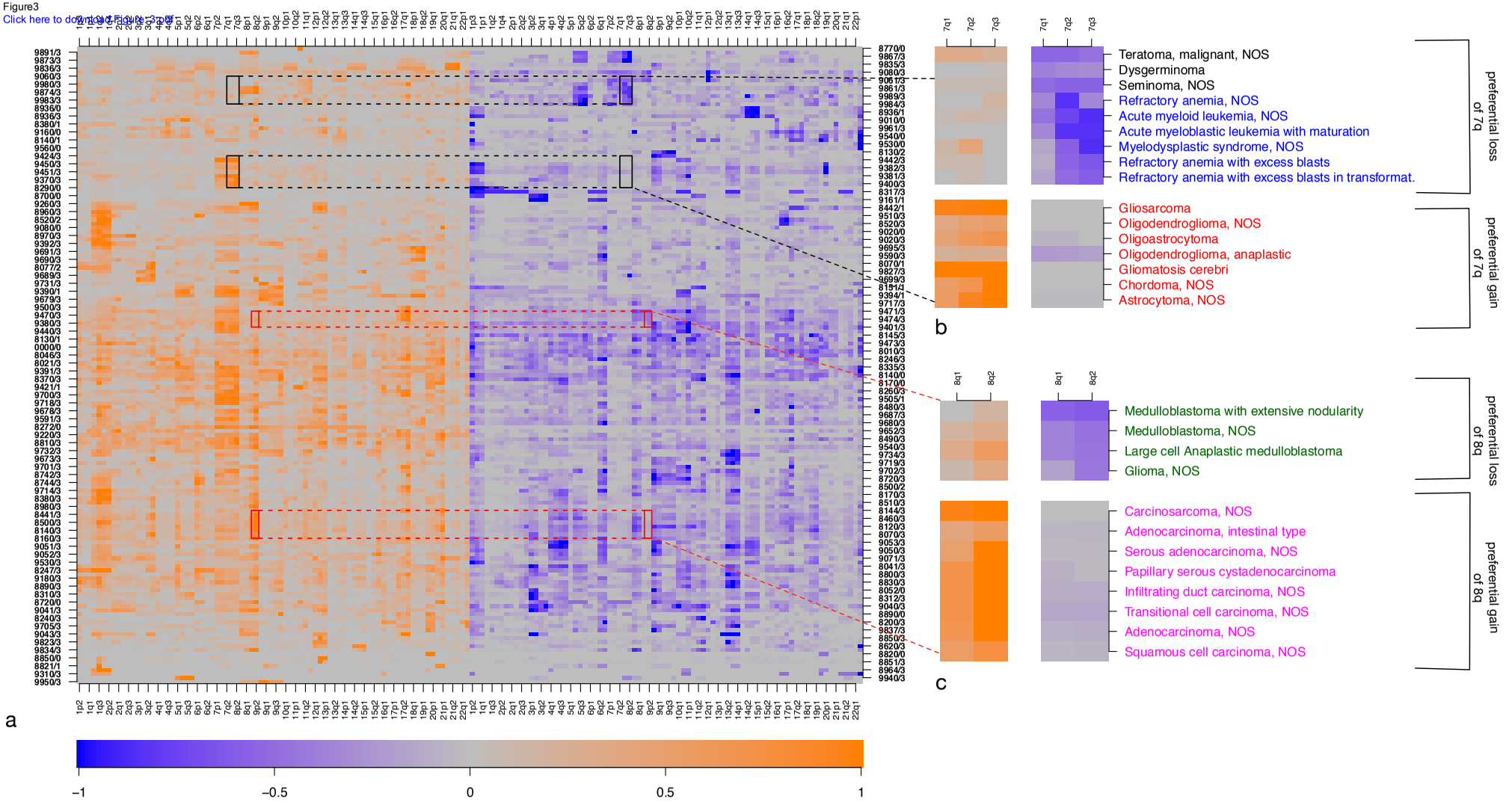
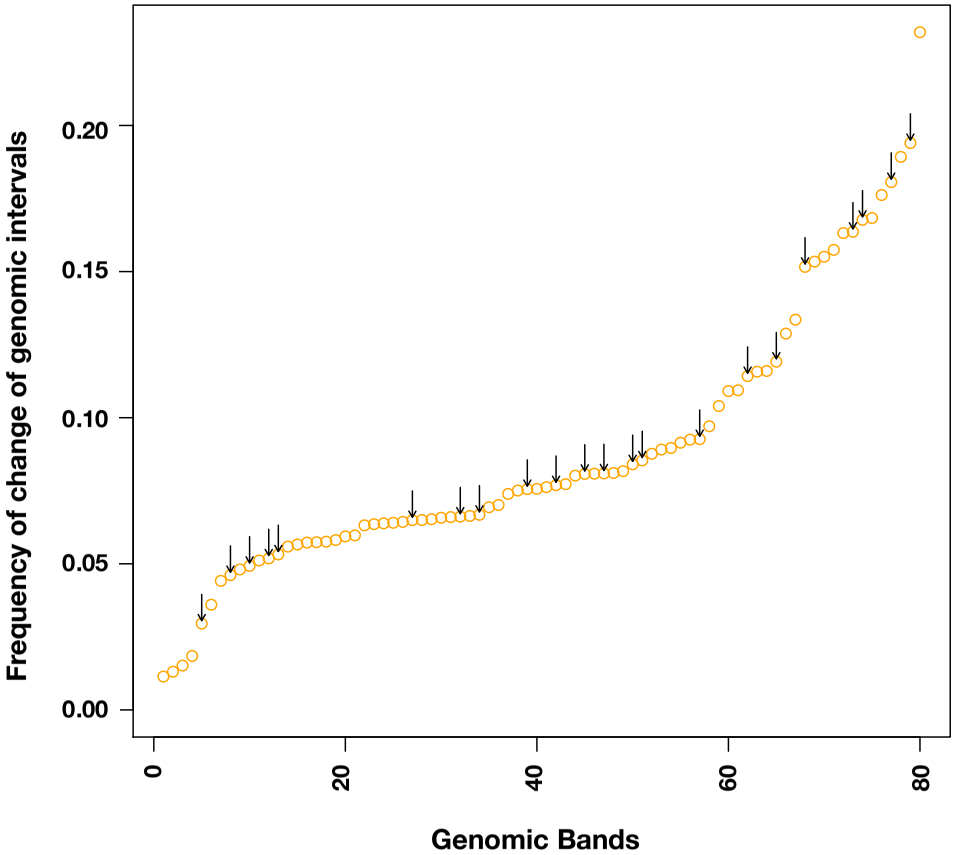
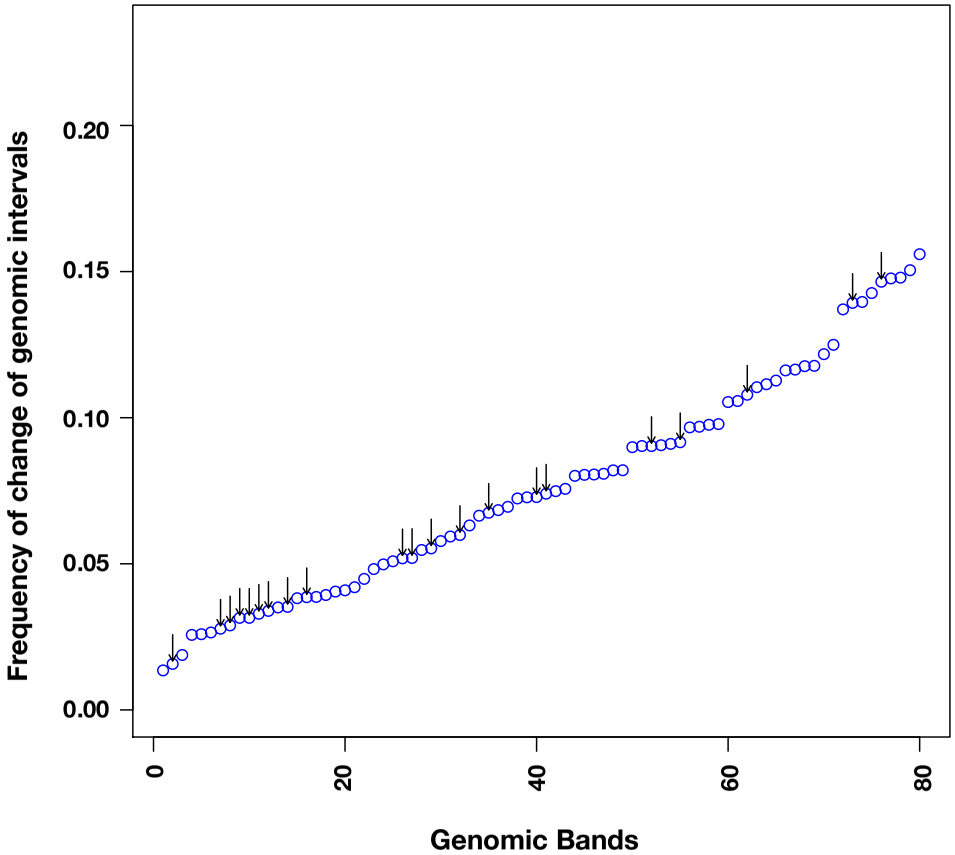


Figure4
Click here to download Figure: 4.pdf

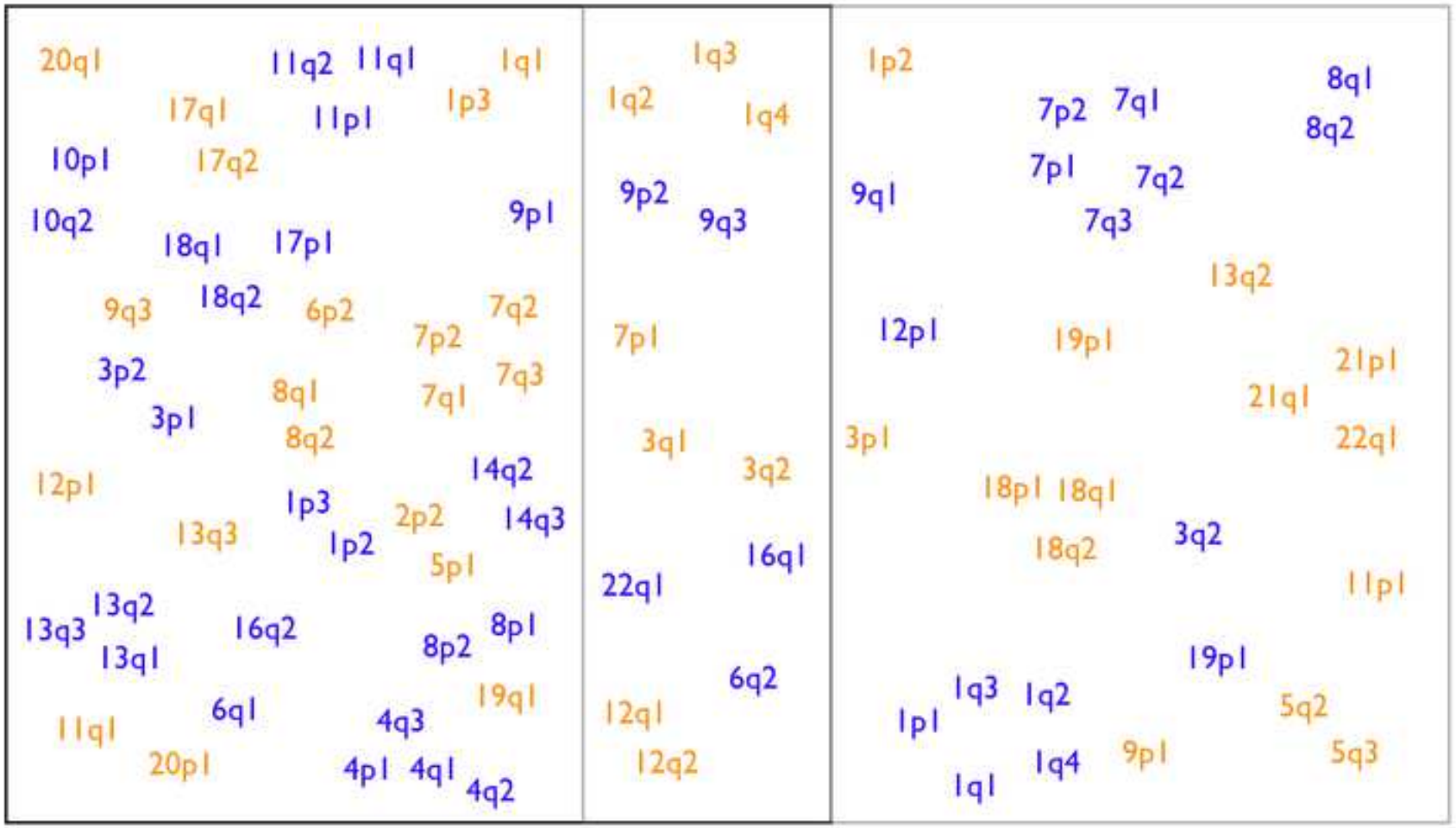


(a) Frequency CNA gains across entire cancer data



(a) Frequency CNA loss across entire cancer data

Figure5
Click here to download high resolution image



10 Acknowledgement

First of all I would like to thank Dr. Michael Baudis for giving me the opportunity to do my Ph.D. thesis under his superstition at Institute of Molecular Life Sciences, University of Zurich. I thank him a lot for always being available next door for any help and active discussion. He has been an amazing supervisor to work with.

Next I would like to thank my Ph.D. thesis committee members Prof. Andreas Wagner, Prof. Christian Von Mering, Prof. Niko Beerenwinkel, Prof. Josef Jiricny and Dr. Hubert Rehrauer for helping and supporting me during my Ph.D. work. They have constantly guided and supported me during my entire Ph.D. thesis. They have always given me loads of input and suggestions on my all scientific queries.

The next thank goes to co-Ph.D. student in my lab Haoyang Chi. He has been a great help in scientific discussion during working hours. He has always kept the working atmosphere very smooth and easy.

My masters students Kamal Fartiyal and Saumya Gupta also own a thank for making the lab environment suitable for work. I learnt quite some stuff from them.

The first of my non-professional acknowledgment goes to my family which includes my mom, dad and my brother for constantly supporting me in all my decisions during my life time. They were really supportive during my Ph.D. thesis, which was sometimes not the easiest.

I would like to thank two of my best friends Michael Probst and Payal Arya. I have known Payal from my Bachelors and she 0has constantly guided and supported me till now. Michael has been a closest friend of mine in Switzerland and has constantly supported me during my stay here. They both have always heard my issues, either personal or profession and have given me good suggestions.

Some other friends and co-workers I would like to thank are Dr. Archana Ayyagari (graduate teacher), Amit Tiwari, Shreya Paliwal, Dr. Carlo Albert, Sushil Kumar, Nessy John, Christoph Moser, Shagun Raina, Neha Daga, Sebastian Schmidt and Rounak Vyas.

11 Abbreviations

A	Amplification
aCGH	Array comparative genomic hybridization
ALCL	Anaplastic large-cell lymphoma
ALL	Acute lymphocytic leukemia
BAC	Bacterial artificial chromosome
B-CLL	B-cell Lymphocytic leukemia
B-NHL	B-non hodgkin's lymphoma
BL	Burkitt's lymphoma
CDCOCA	Complexity dependence of co-occurring chromosomal aberrations
CGH	Comparative genomic hybridization
CML	Chronic myeloid leukemia
CNA	Copy number alterations/aberrations
D	Large deletion
EBV	Ebstein-Barr virus
F	Frameshift
FISH	Fluorescent in situ hybridization
G-path	Gene based pathway enrichment
HIV	Human immunodeficiency virus
HPV	Human papilloma virus
ICGC	International Cancer Genome Consortium
M-FISH	Multiplex Fluorescence In Situ Hybridization
Mis	Mis-sense mutation
N	Nonsense
NSCLC	Non small cell lung cancer
O	Other
PCR	Polymerase chain reaction

ROMA	Representational oligonucleotide microarray analysis
S	Splice site
S-path	Segment based pathway enrichment
SKY	Spectral karyotyping
SNP	Single nucleotide polymorphism
T	Translocation
T-ALL	T-cell acute lymphoblastic leukemia
T-PLL	T cell prolymphocytic leukaemia
TCGA	The Cancer Genome Atlas
TLS	Tree length statistics
UV	Ultra violet
WHO	World health organization

12 Appendix

12.1 Publication 4: arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies

OPEN ACCESS Freely available online

PLoS one

arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies

Haoyang Cai[9], **Nitin Kumar**[9], **Michael Baudis***

Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Abstract

Background: The delineation of genomic copy number abnormalities (CNAs) from cancer samples has been instrumental for identification of tumor suppressor genes and oncogenes and proven useful for clinical marker detection. An increasing number of projects have mapped CNAs using high-resolution microarray based techniques. So far, no single resource does provide a global collection of readily accessible oncogenomic array data.

Methodology/Principal Findings: We here present arrayMap, a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides a platform for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data. To date, the resource incorporates more than 40,000 arrays in 224 cancer types extracted from several resources, including the NCBI's Gene Expression Omnibus (GEO), EBI's ArrayExpress (AE), The Cancer Genome Atlas (TCGA), publication supplements and direct submissions. For the majority of the included datasets, probe level and integrated visualization facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools.

Conclusions/Significance: To our knowledge, currently no data source provides an extensive collection of high resolution oncogenomic CNA data which readily could be used for genomic feature mining, across a representative range of cancer entities. arrayMap represents our effort for providing a long term platform for oncogenomic CNA data independent of specific platform considerations or specific project dependence. The online database can be accessed at http//www.arraymap.org.

Citation: Cai H, Kumar N, Baudis M (2012) arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies. PLoS ONE 7(5): e36944. doi:10.1371/journal.pone.0036944

Editor: Ying Xu, University of Georgia, United States of America

Received January 10, 2012; **Accepted** April 16, 2012; **Published** May 18, 2012

Funding: HC is supported through a personal grant from the China Scholarship Council. NK and MB had received support through the Krebsliga Schweiz and the University of Zurich's Research Priority Program Systems Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: michael.baudis@imls.uzh.ch

[9] These authors contributed equally to this work.

Introduction

Genomic copy number abnormalities (CNAs) are a relevant feature in the development of basically all forms of human malignancies [1]. Many genomic imbalances are recurrent and display tumor-specific patterns [2,3]. It is believed that these genomic instabilities reveal mutations in tumor suppressor genes and oncogenes which eventually result in a clone of fully malignant cells. Investigation of CNA hot spots (chromosomal loci frequently involved in CNA) has proven to be an effective methodology to identify novel cancer-causing genes [4,5]. On a systems level, CNA data along with expression or somatic mutation data is used to detect pathways altered in cancers and to deduce functional relevance of pathway members [6,7]. Since many CNAs have been attributed to specific tumor types or clinical risk profiles, in some entities copy number profiling is employed to characterize different biological as well as clinical subtypes with implications for treatment and individual prognosis. Subtype-associated CNA regions are used to predict causative genes, furthering understanding of biological differences and leading to discovery of new therapeutic targets [8,9].

Throughout the last two decades, molecular-cytogenetic techniques have been applied to scan genomic copy number profiles in virtually all types of human neoplasias. For whole genome analysis, these techniques predominantly consist of chromosomal and array comparative genomic hybridization (CGH), including CNA detection by cDNA and single nucleotide polymorphism (SNP) arrays [10–12]. While chromosomal CGH has a limited spatial resolution of several megabases, the resolution of recent array based technologies (aCGH) is mainly limited due to cost/benefit evaluations instead of technical obstacles. In this article, we use the terms "array CGH" and "aCGH" for all technical variants of whole genome copy number arrays. This includes e.g. single color arrays for which regional copy number normalization is performed through bioinformatics procedures applied to external references and internal data distribution.

The flood of new insights into structural genomic changes in health and disease has led to an increased interest in genomic data sets in genetic and cancer research. Several systematic studies of CNAs across many cancer types have been performed [13,14].

These efforts attempt a more complete understanding of functional effect of CNAs in the context of cancer.

The exponential increase of high resolution CNA datasets offers new challenges and opportunities for large-scale genomic data mining, data modeling and functional data integration. Several online resources have been developed, focusing on different aspects of data content as well as representation [6,15–19]. An overview of some of the prominent examples is given in Table 1. In principle, these databases facilitate access and utilization of CNA data. However, they are limited to specific aCGH platforms and/or single institutions as well as limited disease categories, or, as in the cases of GEO [15] and Ensembl ArrayExpress [16], mainly serve as raw data repositories. To the best of our knowledge, no single data source does yet provide an extensive collection of high resolution oncogenomic CNA data which readily could be used for genomic feature mining, across a representative range of cancer entities.

Here we present "arrayMap", a web-based reference database for genomic copy number data sets in cancer. We have generated a pipeline to accumulate and process oncogenomic array data into a unified and structured format. The resource incorporates associated histopathological and clinical information where accessible.

So far, arrayMap contains more than 40,000 arrays on 224 cancer types from five main data sources: NCBI GEO, EBI ArrayExpress, The Cancer Genome Atlas, publication supplements and user submitted data. Samples of interest can be browsed, visualized and analyzed via an intuitive interface. Computational tools are provided for biostatistical data analysis such as CNA clustering for case specific or for subset data and basic clinical correlations. arrayMap is publicly available at www.arraymap.org.

Results

Data Content

Our combination of both "top-down" (publication driven) as well as "bottom-up" (array data driven) approaches allowed us to identify a comprehensive set of accessible aCGH based cancer CNA data sets and to estimate the ratio of accessible data of the overall published/deposited data.

As main result of the array data driven approach, we extracted 495 series comprising of 32002 arrays, generated on 237 platforms from NCBI's GEO. Among those, raw data files of approximately 29000 whole genome arrays were suitable for inclusion into our data processing pipeline. When reviewing the content of AE, we found that the majority of AE cancer genome data sets were also submitted to GEO. At the time of writing, 11 datasets including 712 arrays not present in GEO had been processed based on AE specific series. Detailed information on the GEO/AE data sets is provided in Table S1.

The top-down procedure was based on our group's continuous monitoring of cancer related articles utilizing genome copy number screening approaches, as established for our "Progenetix" project (www.progenetix.org; [19]). The census date for the literature based data collection was August 15 2011. At this point, we had identified 931 articles discussing a total of 53213 genomic cancer CNA profiles based on aCGH techniques. Of these, 8728 cases out of 199 articles so far had been extracted from publication related sources (e.g. supplementary data tables) and annotated and made been accessible through Progenetix. This data included cases for which only supervised information but no probe data was available (e.g. author annotated Golden Path or cytogenetic CNA regions). Literature based data sets containing probe specific data or with the respective data presented to us by the authors (640 samples) were included into our arrayMap data processing pipeline.

The data content of arrayMap is summarized in Table 2. Current numbers on the website will include changes based on ongoing annotation efforts (i.e. addition of data sets, removal of low quality arrays).

As a by-product of our data collection and annotation efforts, we are able to provide estimates of content and trends for the platform usage and cancer entity coverage for the majority of published data. According to the assigned ICD-O 3 (International Classification of Diseases for Oncology, 3rd Edition) code and descriptive diagnostic text, breast carcinoma predominates as single largest clinical entity with 6459 arrays. Table S2 presents sample sets in arrayMap classified by ICD-O code.

The most widely available array CGH platforms are either based on large insert clones (BAC/P1 arrays) or based on shorter single-stranded DNA molecules (oligonucleotide arrays), which may or may not include single-nucleotide polymorphism specific probe sequences (SNP arrays). Also, although designed for gene expression profiling, cDNA arrays were used by several laboratories for measuring genomic copy number changes. Although all these platforms are considered suitable for whole genome CNA analysis, their probe densities and other parameters can affect specific features of the analysis results [20–23]. Table S3 lists the general platform types and corresponding overall numbers of the data registered in arrayMap.

Table 1. Prominent online resources of genomic data.

Name	Address	Platform(s)	Data format	Comment
GEO [15]	www.ncbi.nlm.nih.gov/geo	263	raw & normalized probe signal intensity	largest microarray data repository
ArrayExpress* [16]	www.ebi.ac.uk/arrayexpress	16	raw & normalized probe signal intensity	many duplicate data in GEO
TCGA [6]	cancergenome.nih.gov	1	segmentation data	raw probe data is limited to download
CanGEM** [17]	www.cangem.org	38	normalized probe signal intensity	including many types of microarray data
CaSNP [18]	cistrome.dfci.harvard.edu/CaSNP	8	average copy number & graphic	focus on SNP array data
Progenetix [19]	www.progenetix.org	235	ISCN*** & golden path	data from publications

Data up to 29 April, 2011.
*excluding data both in GEO and ArrayExpress.
**statistical information only including CGH, SNP and cDNA data.
***International system for human cytogenetic nomenclature.
doi:10.1371/journal.pone.0036944.t001

Table 2. aCGH data integrated in arrayMap.

Data Source	Arrays	Cases	Series	Platforms	Publications
GEO	32002	25728	495	237	490
ArrayExpress	712		11	16	11
TCGA	7249	3594	19	1	*
Publication Supplements	>4578**	4578			137
Author Submission	556	539	8	7	

Data up to 29 April, 2011.
*Due to lack of publication information, there may be a small amount of duplicate data in GEO.
**Array number may be higher than case number since reported results per case occasionally may be based on more than one array. The number does not include data presented both in publication supplements as well as GEO.
doi:10.1371/journal.pone.0036944.t002

In reviewing the technical platform composition, two related trends become apparent (Figure 1). Originally developed in groups with expertise in molecular cytogenetics and cancer genome analysis, printed large insert clone arrays (BAC/P1) were the first whole genome CNA screening tools with a spatial resolution surpassing that of chromosomal CGH. Other groups re-employed cDNA arrays, developed for expression screening, for genomic hybridizations. However, over the last years one can observe the overwhelming use of various industrially produced oligonucleotide array platforms, which compensate their low single probe fidelity through a probe density at 1–3 orders of magnitude higher than common for BAC/P1 arrays. Another reason for the success of oligonucleotide arrays is the integration of SNP specific probes, which in principle allows to use of the same experiments for genetic association studies and the evaluation of copy number neutral loss of heterozygosity regions [12,24,25].

Data Access and Usage Scenarios

Based on our experience from the Progenetix project, a strong emphasis was put on a user friendly data interface. Here, we followed a ''dual user type'' scenario: Users without bioinformatics background should be able to intuitively visualize core data features as well as to perform standard analysis procedures, while for bioinformaticians the formatted database content should be accessible to use with their analysis tools of choice.

Query interface. Data browsing in arrayMap is based on two types of query methods: search by experimental series metadata and search by sample features.

In the series query form, users can perform various search options by specifying (i) descriptive diagnosis text; (ii) disease classification (ICD-O 3 code(s)); (iii) disease locus (ICD topography code(s)); (iv) PubMed ID; (v) technique(s); (vi) series ID. For sample specific queries, additional features are available: sample ID; platform ID or description; and single or combined regional CNAs. Users can input gene name(s) in ''regional CAN'' search field. When at least two characters are entered into the field, suggestions based on a HUGO gene list are displayed for selection. Gene selections will be converted to genomic locations.

In the results table, associated array information is displayed. A number of links to additional and/or outside data is provided, according to the information available: the corresponding PubMed entries; the original GEO/AE accession display page for more complete information; the case and publication entries on the Progenetix website for further analysis; and importantly the array specific data visualization page.

Data download options. On pages resulting from sample queries or sample data processing, users are presented with options to download sample data based on the current queryÕs return. Currently, three different file types are offered: JSON files, tab separated feature files and segments list files. These files enable bioinformaticians to perform further analyses based on their tools of choice. Particularly, the JSON format can be used for direct database import (e.g. MongoDB) or can be deparsed by common libraries (e.g. JSON.pm), or being read into web applications.

Array probe data visualization. In the array plot interface, original plots of genomic array data sets can be searched and visualized (Figure S1). Default threshold parameters which were either provided with the data or assigned during the initial visualization will be loaded. In single array visualization, the general view of probe distribution and post-thresholding segmentation results are displayed for the whole genome as well as for each individual chromosome. If multiple arrays are retrieved, users can select sample data for downstream analysis procedures. Figure S2 shows the screenshot of single array visualization.

Users can segment the raw data values and re-plot the results after revising the following parameters:

- Golden path edition, default HG18/NCBI Build 36. This is still the commonly used version of the human reference genome assembly. At the moment, coordinates of probes from all platforms were remapped to HG18. For the near future, we intend to allow for a selection of updated genome editions.
- Chromosomes to plot, default 1 to 22. Single or all chromosomes can be selected for re-plotting. To avoid gender bias, most platforms do not contain probes in chromosome X and Y during the design.
- Loss/gain thresholds. Cut-offs from which a segment is considered a genomic loss or gain. The optimum thresholds may vary between platforms.
- Region size in kb. Sets a filter to remove CNA below (e.g. probable noise) or above (e.g. exclude non-focal CNA) a certain size range.
- Minimal probe numbers for segments. This parameter can be used to limit the minimal number of probes required for a segment to be considered (e.g. to remove aberrant segmentation due to probe level noise). Empirical examples would be values of 2–3 for high quality BAC arrays and 6–10 for Affymetrix SNP 6 arrays.
- Plot region. Single genomic region to be plotted, overriding the chromosome selection above. When selected, plots with this region will be generated for all current arrays. This is valuable to e.g. display the CNA status and copy number transition points for specific genes of interest (Figure S3).

Zoom-in visualization of focal CNA. Figure 2 shows the visualization of focal genomic imbalances, e.g. to identify genes of interest targeted by focal CNA. The whole genome view of GSM535547 (human high grade glioma sample analyzed by Agilent Human Genome CGH Microarray 244A) shows a small regional deletion in chromosome 9p21. When plotting the approximate locus of the deletion (specified as chr9:21600000-22400000), genes, probes and chromosome bands in this zoomed in region are shown. Two genes, MTAP and CDKN2A can be seen as being localized in a potential homozygously deleted region. The focal deletion of these known tumor suppressor genes [26,27] points to their specific involvement in the glioblastoma sample analyzed here.

Querying compound CNA. The concept of focal CNA detection can be integrated with a global search for arrays

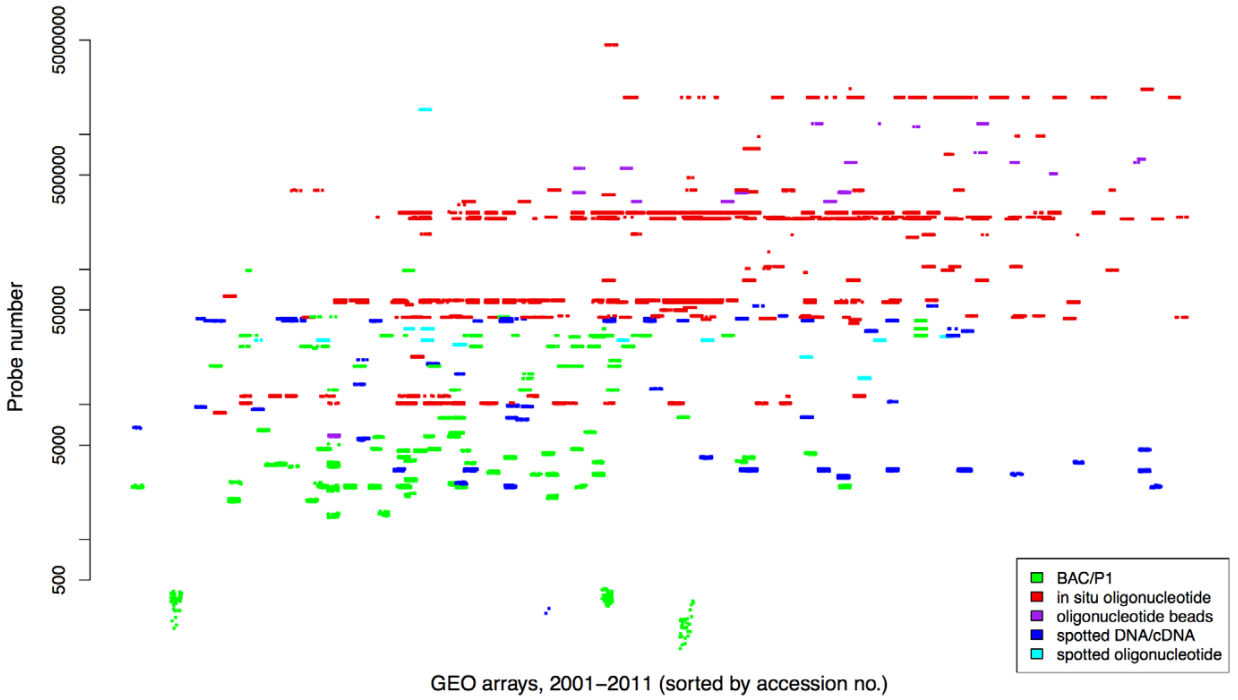


Figure 1. Distribution of resolutions and techniques of GEO platforms. Each point represents a genomic array. The Y axis is labeled with probe number in log scale. The X axis denotes the time sequence of array data generation. From left to right are years from 2001 to 2011.
doi:10.1371/journal.pone.0036944.g001

containing gene specific regional imbalances. As an example, we demonstrate the search for arrays displaying imbalances in 4 gene loci associated with glioblastoma: EGFR, a transmembrane receptor and proto-oncogene [28]; PTEN, a tumor suppressor gene [29]; ASPM, frequently overexpressed in glioblastoma relative to normal brain tissue [30]; and CDKN2A (see above). In the "Search Samples" form, the "Match (Multiptle) Regions & Types" can be used to specify the genomic regions of those four genes including the expected CNA type: for EGFR (chr7:55054219-55242524:1), PTEN (chr10:89613175-89718511:-1), ASPM (chr1:195319885-195382287:1) and CDKN2A (chr9:21957751-21984490:-1), respectively. When executing the query, these regions were matched with the whole database and returned cases which have imbalances overlapping all these regions. When excluding controls and "worst quality" datasets, 303 out of 42421 arrays could be identified matching all four CNA regions. In addition to glioblastoma, several other types of cancer cases were among the results, including e.g. neuroblastomas, breast carcinomas, melanomas and lung carcinomas, which is in accordance with some previous observations [31–34]. CNA and associated data of those cases can be processed by online tools for further analysis and visualization (Figure S4) or downloaded for offline processing.

Copy number profiling of selected cancer entities. One aim of arrayMap is to allow researchers to conveniently perform aCGH meta-analysis across different platforms. By selecting a single or several cancer entities e.g. based on their ICD entity codes or diagnostic keywords, users are able to generate disease specific CNA frequency profiles or to compare profiles of different cancer types.

As an example, we used ICD-O code 9440/3 (glioblastoma, NOS) to query the database. 1478 arrays from 25 publications were returned and passed to our suite of online analysis tools. Chromosomal ideograms and histograms were generated representing the frequency of copy number aberrations identified over the whole dataset (Figure 3A). In the overall aberration profile, the most common genomic imbalances included whole chromosome 7 gain and chromosome 10 loss, as well as focal gains e.g. on bands 1q21 and 17q21. In our example dataset, a prominent focal deletion hot-spot was centered around 9p21.3 (921 of 1478 arrays, 62.31%) which had been discussed previously [35]. The distribution of CNAs over the individual arrays was visualized through a matrix plot (Figure 3B). As additional information to the frequency histograms, this form of visualization facilitates e.g. the detection of CNA patterns among individual arrays as well as the concordance of individual CNAs (e.g. here the arm-level changes in chromosome 7 and 10).

In the matrix plot, clicking on a certain segment would open the related view in the UCSC genome browser [36], for detailed information related to this genomic region (SVG plot only). The plot order of arrays can be re-sorted according to ICD morphology, ICD topography, clinical group or PubMed ID, which can be helpful in associating CNA patterns to external classification categories. For the selected classification criterium (default: ICD morphology), regional CNA frequencies for cases matching the different values will be visualized through a heatmap (Figure 3C); this feature is especially useful when comparing a number of different primary classification criteria.

An Overall Genomic Copy Number Profile of Cancer

Our high quality core dataset in arrayMap was used to generate an overall cancer copy number aberration profile based on 29,137 arrays (Figure 4). This data represented 177 cancer types according to ICD-O 3 code, with 59 types among them contained

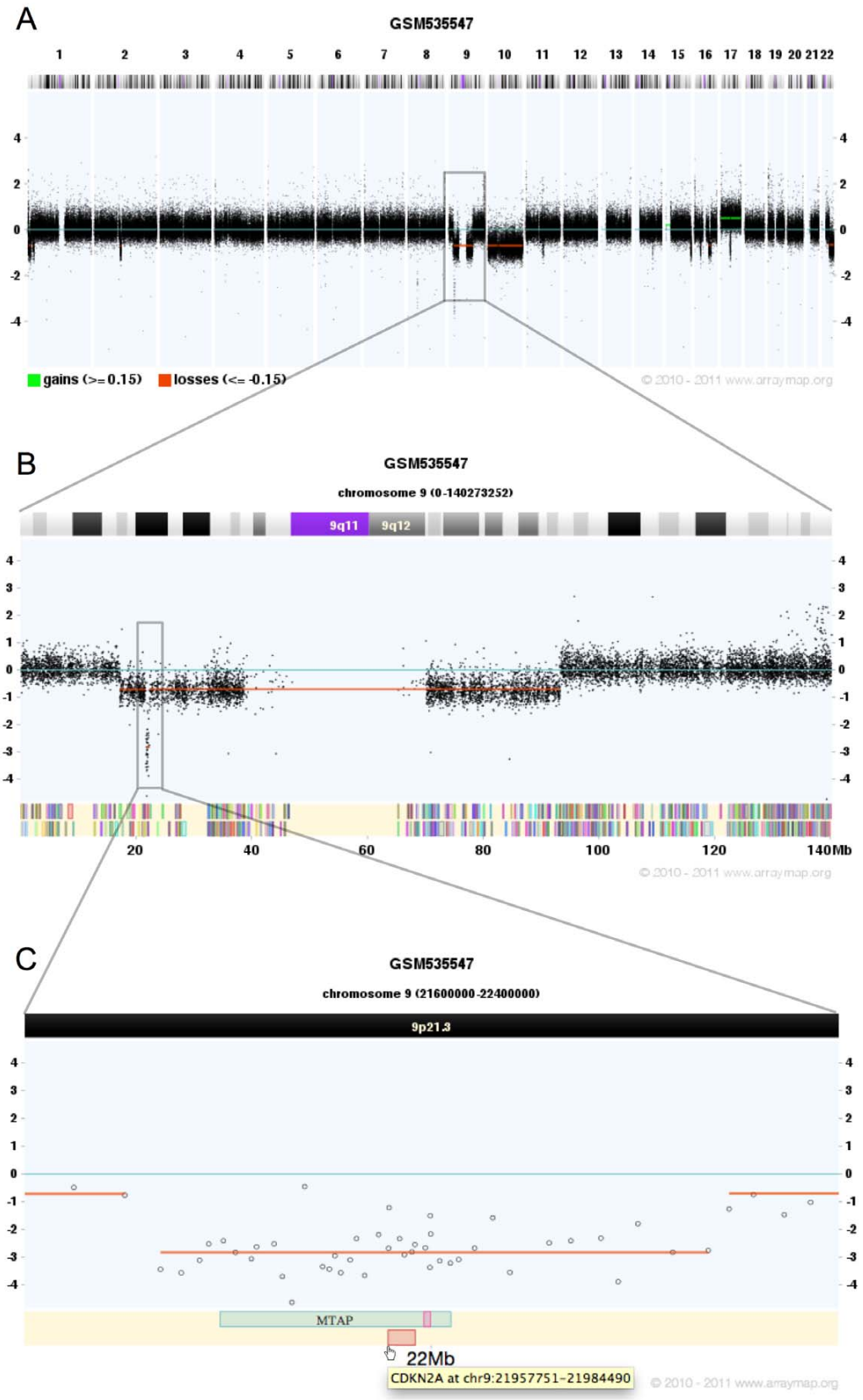

A
GSM535547
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
gains (>= 0.15)
losses (<= -0.15)
© 2010 - 2011 www.arraymap.org
B
GSM535547
chromosome 9 (0-140273252)
9q11
9q12
20 40 60 80 100 120 140Mb
© 2010 - 2011 www.arraymap.org
C
GSM535547
chromosome 9 (21600000-22400000)
9p21.3
MTAP
22Mb
CDKN2A at chr9:21957751-21984490
© 2010 - 2011 www.arraymap.org

Figure 2. Zoom-in visualization of focal CNA. (A) GSM535547 (human high grade glioma, Agilent CGH 244A) shows high quality of probe hybridization signal. CNAs are easy to distinguish. (B) When zoom-in the whole chromosome 9, an approximately 80 MB deletion is displayed, with two breakpoints located in p and q arm respectively. In addition, a small regional deletion in 9p21 is quite clear. Color bars in lower region of the panel represent 848 genes located in chromosome 9. (C) Zoom in the potential homozygously deleted region in 9p21 by specifying the exact region: chr9:21600000-22400000. The zoomed-in plot shows probes, chromosome band and two tumor suppressor genes, MTAP and CDKN2A. Gene name and location will be given while mouse hover. They link to UCSC genome browser with additional information.
doi:10.1371/journal.pone.0036944.g002

more than 50 arrays. Overall, one of the most common genomic alteration is copy-number gain of chromosome band 8q24, which is found in 30% of total samples. According to the COSMIC [37] database, the most significant cancer gene in this region is MYC. It is a well-documented oncogene codes for a transcription factor that is believed to regulate the expression of 15% of all genes, including genes involved in cell division, growth, and apoptosis [38,39]. Other common imbalances observed in at least 25% of oncogenomic arrays included gains of regions on e.g. 17q21 (29%), 1q21 (33%) and loss of regions on e.g. 8p23 (32%) and 9p21 (25%), including focal deletions of the CDKN2A/B locus (Figure 2).

While the overall CNA frequency distribution points towards DNA features targeted in multiple entities, this information is insufficient for deriving molecular mechanisms associated with specific cancer types. The genomic heterogeneity of different neoplasias is reflected in the varying patterns of regional CNA frequencies. Based on our core dataset, we have generated a heatmap-style visualization of frequency profiles for all ICD-O entities containing more than 50 arrays (Figure S5). The striking

Figure 3. Copy number profiling of glioblastoma. (A) Chromosomal ideogram and histogram showing frequency of copy number aberrations. Percentage values corresponding to gains (yellow) and losses (blue) identified over the whole dataset. The most frequent imbalances include gain of chromosome 7 and loss of chromosome 10, 9p21.3. (B) Matrix plot of 1478 glioblastoma cases. The Y axis represents individual samples. The distribution of genomic copy number imbalances reveals the individual aberration patterns of glioblastoma. (C) Heatmap of regional CNA frequencies for 1478 arrays. The intensity of green and red color components correlates to the relative gain and loss frequencies, respectively. If dataset contains cancer subtypes, cancers with similar CNA frequency profiles will be clustered together, such that differences between subtypes will be revealed (e.g. see Figure S4H).
doi:10.1371/journal.pone.0036944.g003

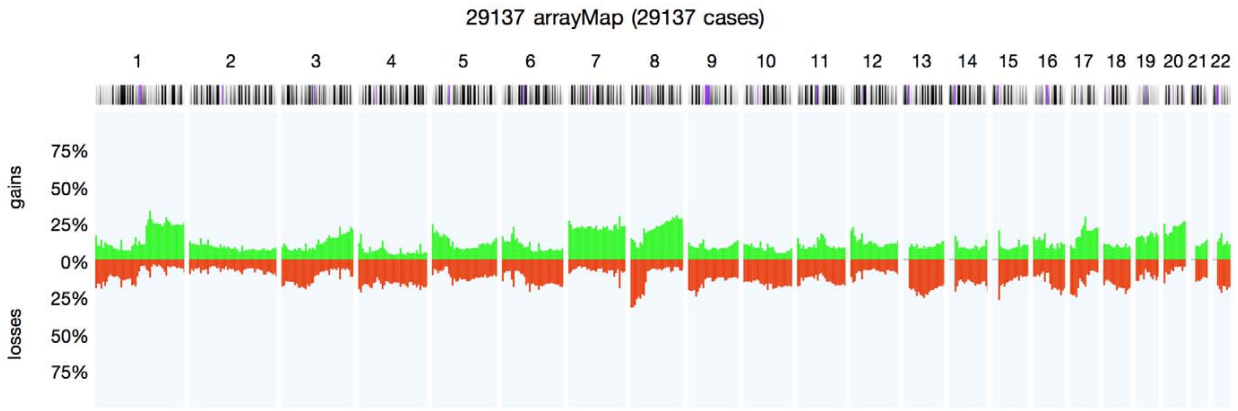


Figure 4. The overall cancer copy number aberration profile consisted of 29137 arrays. This plot represents 177 cancer types according to ICD-O 3 code. Percentage values in Y axis corresponding to numbers of gains (green) and losses (red) account for the whole dataset.
doi:10.1371/journal.pone.0036944.g004

patterning of the CNA profiles indicates the non-random occurrence of CNAs, and should be seen as an invitation to explore e.g. CNA similarities shared by separate histopathological entities, as a way to transpose knowledge about pathophysiological mechanisms.

Discussion

arrayMap was developed to facilitate the progress of oncogenomic research. Our aim is to provide high-quality genomic copy number profiles of human tumors, along with a set of tools for accessing and analyzing CNA data. The service has been implemented with a straightforward web interface, including search options for CNA features and clinical annotation data. All assembled datasets are processed into platform independent segmentation and, for the vast majority of arrays, probe level data files, and are presented in consistent formats. Importantly, the direct access to precomputed probe level data plots supports a rapid evaluation of experiments for features of interest. As a curated database using standardized annotation schemes (e.g. ICD classification), arrayMap facilitates the exploration of cancer type specific CNA data, as well as the statistical association of genomic features to clinical parameters.

arrayMap is a dynamic database that is being continuously expanded and improved. We will review the existing and newly published articles to update the database periodically. Over the past decade, we have witnessed a rapidly increasing number of aCGH publications, which gives us sufficient evidences to anticipate that cases in our database will continue to be deposited at a high rate. Although arrayMap is not a user driven repository, we welcome and support users interested in using the site for yet undisclosed data, if they agree on data sharing upon publication.

Although, in contrast to the continuous data from expression analysis, copy number analysis explores discrete value spaces (countable number of DNA copies, for segments defined by genomic base positions), interpretation of the data can vary due to different low level (e.g. signal/background correction) and higher level (e.g. segmentation algorithms, regional or size based filtering) procedures. In that respect, we have to emphasize that the results of our data processing and annotation procedures are open to scrutiny. We encourage a critical review of individual results, and are open for suggestions regarding improved processing procedures for specific platforms.

In this paper, we have provided example scenarios of using arrayMap on different levels, i.e. locus centric and for entity profiling. We believe that systematic analyses will help researchers to discover features which are indiscernible in individual studies, and thus bring new insights for understanding of disease pathology and the development of new therapeutic approaches [40–43]. We expect that researchers will integrate arrayMap data with their own analysis efforts, e.g. to increase sample size or for result verification purposes. We hope that this database will promote further evolution of microarray data meta-analysis. ArrayMap provides access to more than 200 tumor types, which makes it suitable for research across cancer entities. Furthermore, normal sample controls are of vital importance for genomic imbalances studies. ArrayMap includes more than 3000 normal samples from healthy individuals or from normal tissues of cancer patients. These data could be integrated as reference dataset e.g. to account for copy number variation data superimposed on the tumor profiling results.

In the near future, with the continuous accumulation of very high resolution CNA data from genomic arrays and next-generation sequencing experiments, it will become possible to integrate these data into systems biology methods to elucidate effects of genomic instability, and describe the results from more perspectives. Envisioned examples would be e.g. the identification of genes that are involved in metastasis and treatment response; identification of chromosomal breakpoints distribution in cancer; and modeling functional networks in cancer by systems biology approaches.

Methods

Dataset Collection

Raw experimental data from a variety of platforms and repositories were extracted. They were converted to an uniform format which is suited to our reanalysis and visualization system. After a series of parsing procedures, the called copy number data is stored in arrayMap. The flowchart of arrayMap data collection and analysis is as shown in Figure 5. Five main data sources are integrated into arrayMap:

GEO/AE. For extracting appropriate data Series from GEO/AE, two basic criteria have to be fulfilled. First, the raw data has to be from human malignancies analyzed by BAC, cDNA, aCGH or oligonucleotide arrays. Second, the array platform must be genome wide, with the optional omission of the sex chromosomes.

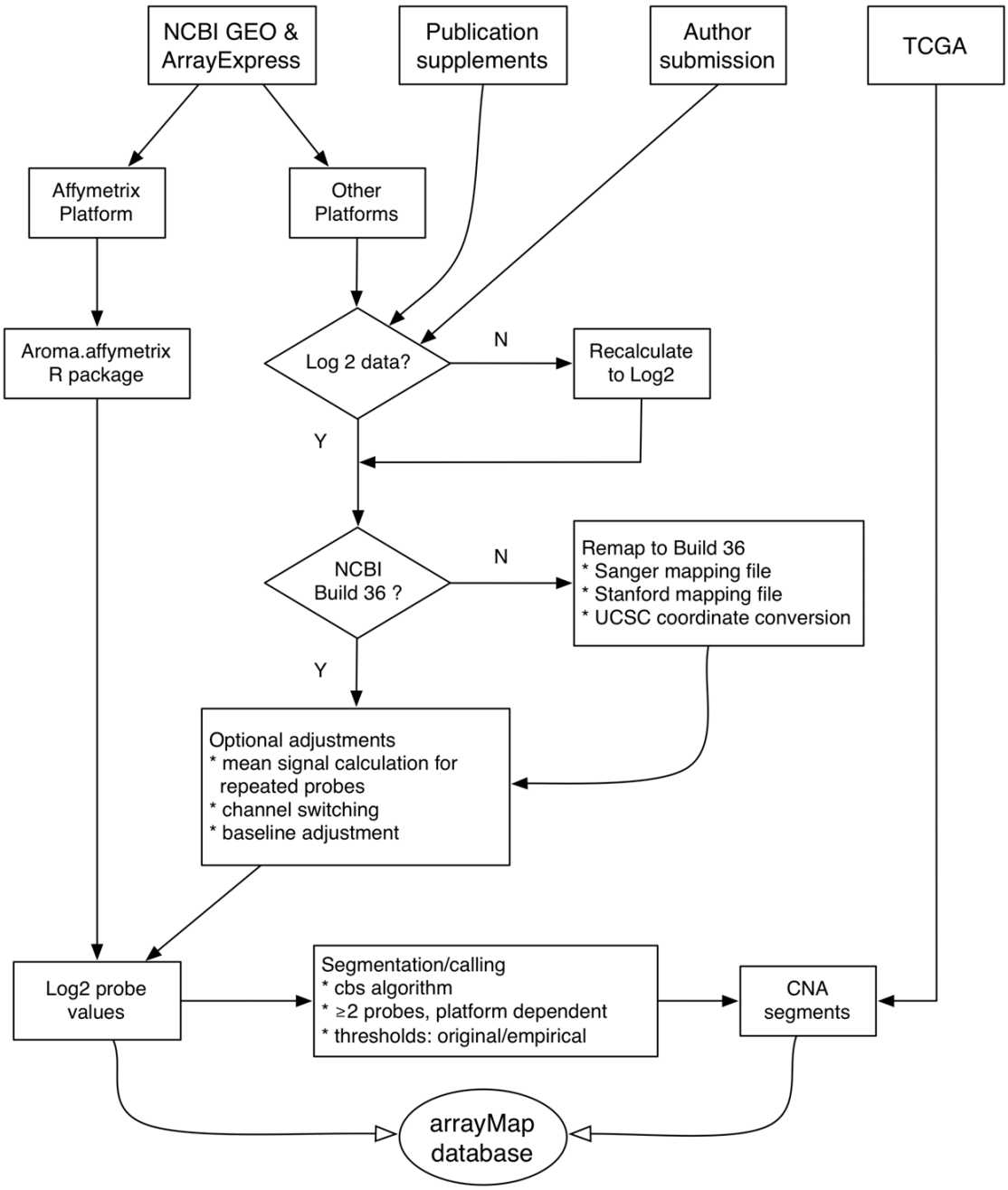


Figure 5. The flowchart of arrayMap data collection and analysis procedures. Publicly available raw data or segmented data was collected from the respective data sources. Files were re-processed by distinct procedures, according to the different data types. Probe coordinates were remapped to the most commonly encountered human reference genome assembly (NCBI Build 36/hg18). All probe specific ratios were converted to log2 values. Thresholds for genomic gain and loss were obtained from the original publications or series annotations; if not available, empirical thresholds were assigned. A minimum of 2 probes was required for calling a CNA segment, with higher values used on high-density arrays and/or in cases of excessive probe level noise. Processed probe and segment information was converted to uniform formats and stored in per-sample text files, which are accessed through the arrayMap web applications.
doi:10.1371/journal.pone.0036944.g005

Chromosome or region specific arrays were excluded because they were not able to reveal the whole genomic profile of the respective cancer. Associated clinical data was extracted if available.

TCGA. Segmentation data with available clinical information was extracted and incorporated into the database. Due to data sharing restrictions, TCGA data is an exception in that, so far no probe level data is incorporated into arrayMap. This exception

was accepted since users will be able to access individual TCGA datasets through the projects web portal at http://tcga-data.nci.nih.gov/tcga/.

Publications. Many aCGH datasets can be found in the text or supplementary files of publications. In order to collect data from publications, we relied on our Progenetix projectÕs setup. Data in Progenetix is manually curated. The collection strategies are:

- literature mining using complex search parameters through PubMed
- identification of called aCGH data, in GP annotation or tabular format (article, supplementary tables)
- evaluation of supplementary files for probe specific data tables
- follow-up on article links outs, to repository entries or referenced datasets

User submission. User submitted data was provided in a number of formats which were converted to the standard format as described. Although we accept and support private datasets, we insist on integration of at least the genomic and core clinical data (e.g. disease classifiers) upon publication of the datasets analysis results.

Dataset Analysis

Probe remapping. A pipeline has been generated for determining the genomic positions for the tens to hundreds of thousands array probes with reference to a common genome Golden Path edition. For each array platform, the genome positions of probes were remapped to the current commonly used version of the human reference genome assembly (NCBI Build 36.1/hg18). Specific mapping procedures were employed for different types of probes. BAC clones were firstly remapped according to the clone sets information of Sanger/DECIPHER database [44]. If the probe position was not available, the UCSC Genome annotation database [36] (release hg18) was used for compensation. After these two steps, a mean of 98% of the BAC clones were remapped. For IMAGE clone sets, only the UCSC Genome annotation database was used. The average remapping rate of IMAGE clones was 91%. Affymetrix raw CEL data files were analyzed based on hg18 library files, namely the output segments have hg18 coordinates. The summary of the percentage of mapped probes is given in Table 3. The mapping details for each platform can be found in the (Table S4).

Probe signal normalization. The array data available was given in a variety of formats, most frequently as log2 ratio of probe hybridization intensity. In order to make data from different platforms directly comparable, all other types of normalized values were converted to log2. For dye swap experiments, reference/tumor intensity ratios data was "reversed" representing a tumor/reference value. For some two-color arrays for which only raw signal intensity were provided, the normalized log2 ratio for each probe was calculated by.

$$r = \log 2((T_s - T_b)/R_s - R_b).$$

where T_s and T_b represent tumor sample intensity and tumor channel background intensity respectively, and R_s and R_b represent reference sample intensity and reference channel background intensity respectively. If multiple instances of the same clone exist, the average signal intensity of the certain clone was considered.

To call gains and losses according to normalized log2 ratio is an important step to identify copy number imbalances. For each re-analyzable dataset, related publications were explored to obtain original threshold descriptions. If this information was not available, empirical thresholds were assigned and resulting CNA calls were visually compared with probe value plots. Processing method and threshold information for each array are provided in the Table S5.

Affymetrix genotyping arrays. For the widely used Affymetrix GenomeWide SNP arrays, raw CEL files were downloaded and underwent a massive re-analysis using the R package aroma.affymetrix [45] with the CRMAv.2 method [46]. During the processing step, approximately 50 normal sample arrays were employed as a reference set for each array type to reduce the noise level. Normal tissue arrays from different labs were extracted and used to build the reference dataset. In order to obtain high quality arrays, we excluded arrays which contain segments greater than 3 mega-bases, since copy number variations are always smaller than 3 mega-bases. The list of normal tissue reference arrays is giving in Table S6.

Quality control. In our review of array data deposited in GEO or collected from publication supplements we encountered a large number of individual data sets with insufficient or limited probe quality. Also, for samples of unprocessed raw data (e.g. Affymetrix CEL files), we found that QC measures reported previously (e.g. call rate [47], NUSE [48], RLE [48]) only had a limited accuracy for detection of arrays with inadequate probe level data. Currently, the most viable strategy for quality assessment of processed, heterogeneous copy number arrays is the visual inspection of probe plotting and segmentation results through an experienced researcher. For the first arrayMap edition we generated a quality classification system, which contains a total of 4 categories based on inspections of genome-wide array plots:

- Excellent. Probe signal distribution is significantly different between normal regions and imbalance regions. Signal baseline is distinct and unique, making segmentation threshold realistic appearing. Chromosomal changes are pretty clear.
- Good. In general good quality. Probe signal may contain some noise, but tolerable. Chromosomal changes are distinguishable.

Table 3. Percentage of remapped probes according to platform types.

Platform type	Average mapping rate	Number of arrays	Number of GPLs
Original HG18 (Build 36)	NA	1583	40
in situ oligonucleotide	99%	21678	55
BAC/P1	98%	5464	55
spotted DNA/cDNA	91%	2365	82

doi:10.1371/journal.pone.0036944.t003

- Hypersegmented. Serrated distribution of probe signal intensities, causing dozens of separate peaks and discontinuous segments. Chromosomal changes are always up to several hundreds and smaller than 5 mega-bases.
- Noisy. Probe signal intensities are highly scattered, but well-distributed, with high standard deviation, resulting in the inability to differentiate copy number changes.

Depending on the intended research purpose this basic classification system can be used for a pre-analysis triage of copy number data. Applying stringent review criteria we identified a core dataset with "excellent" quality arrays accounting for approximately 60 percent of total arrays. We are currently working on a platform independent quality assessment system for genomic arrays, which will be implemented in future versions of the arrayMap resource.

Associated data. For arrayMap, data is stored with separate datasets for each array. This is in contrast to the Progenetix database, for which technical replicates where available are combined into case specific CNA profiles. In arrayMap, technical replicates are assigned an identical case identifier to facilitate downstream statistical procedures including e.g. clinical data correlations. The assignment of the correct diagnostic entity to each sample is an essential step in generating a binding between genomic and associated data points. At the same time, to ensure annotation consistency and make the retrieval process more efficient, for all CNA profiles the following data points were manually collected from GEO/ArrayExpress and published papers if available.

- Descriptive diagnostic text, as available through the original source
- Diagnostic classification according to the International Classification of Diseases in Oncology (ICDO 3, morphology with code)
- Tumor locus according to ICD (ICD topography with code)
- Source of material (e.g. primary tumor, cell line, metastasis)
- Clinical parameters where available, including age, gender, grade, clinical stage (TNM coded), recurrence/progression, time to recurrence/progression, death and followup

Web Server. An online interface of arrayMap database was created using Perl common gateway interface (CGI) and R scripts running on Mac OS X Server. Sample and series data is stored using a MongoDB database eingine (http://www.mongodb.org). Precomputed array plots are stored as flat files, mostly in both SVG and PNG versions. The online release of the service has been optimized to be compatible with major browsers supporting current web standards (CSS2, HTML5, XML with inline SVG; e.g. Safari $>=3.0$, Firefox $>=3.0$, InternetExplorer $>=9$, Google Chrome) with limited fallback support. Dynamic graphics provided in the array plot module were implemented as server side services by technologies including XML/XHTML, JavaScript, SVG and HTML5 Canvas.

For the future, we intend a quarterly database content revision to ensure inclusion of newly published articles and GEO/AE entries. Archived versions of the sample annotations will be made available upon special request. Additional feature and small data updates will be performed as seen necessary. The "News" page of Progenetix/arrayMap will be used for feature and content announcements.

Supporting Information

Figure S1 Array data sets visualization. Original plots and optimized parameters for GSE21530 which contains 8 intimal sarcoma samples hybridized on Agilent CGH Microarray 244A platform. The normalized probe signal log2 ratios and post-thresholding segmentation results for each array are intuitively displayed. Genomic alterations are represented by horizontal green (gain) and red (loss) lines. Alterations defined here as regions with log2 ratio >0.15 or <-0.15. Simplified schemas of CNAs link to UCSC genome browser for further review.
(PDF)

Figure S2 Screenshot of single array visualization. ArrayMap plots for GSM630977 (acute myelogenous leukemia). Besides the whole genome view, subviews of each chromosome are displayed as well. From these plots, different kinds of genetic variation events are clearly revealed, e.g. massive genomic rearrangement in chromosome 6; arm-level gain of chromosome 8q and 3MB focal change around 1p31.3. Through the "Plot Array Data" interface, users can segment the raw data values and re-plot the results with customized parameters.
(PDF)

Figure S3 Plot single genomic region. In the "Plot Array Data" interface, input the precise location (chr5:1100000-1400000) in "Plot Region" field. Plots with this region were generated for all 8 arrays in the current series (GSE21530). In this region, there are 5 genes which are shown schematically as colored boxes. CNA status and copy number transition points for these genes are displayed.
(PDF)

Figure S4 Compound CNA query. (A) Four gene loci associated with glioblastoma (EGFR, PTEN, ASPM and CDKN2A) were inserted into "Match (Multiple) Regions & Types" field. 303 out of 42421 arrays were returned. (B) Classification information of these 303 arrays were displayed and can be selected for the following analysis. (C) Statistical and plot parameters can be customized. Associated data was processed by online tools, and returned results included: (D) Chromosomal ideogram and (E) histogram, show frequency of copy number aberrations; (F) Matrix plot reveals the aberration pattern of selected arrays; (G) Array classification tree generated by hierarchical Ward clustering, arrays with similar frequency of CNA are part of the tree branch. (H) Heatmap of CNA frequencies clustered by clinical group.
(PDF)

Figure S5 Heatmap of frequency profiles for 59 cancer types. Heatmap visualization of frequency profiles for all ICD-O entities containing more than 50 arrays in our core dataset. Region specific gain/loss frequencies were mapped to 1MB intervals. The intensity of colors (green: gains; losses: red) corresponds to the relative frequency of CNAs for each interval.
(PDF)

Table S1 Entities extracted from NCBI GEO and EBI ArrayExpress.
(XLS)

Table S2 Cancer entities grouped by ICD-O code.
(XLS)

Table S3 Platform type distribution in arrayMap.
(XLS)

Table S4 Probe remapping rate for platforms.
(XLS)

Table S5 Processing method and threshold for calling genomic gains and losses.
(XLS)

Table S6 Normal tissue reference arrays for Affymetrix platforms.
(XLS)

Acknowledgments

We want to thank Christian von Mering, Homayoun Bagheri, Henrik Bengtsson and Nuria Lopez-Bigas for helpful discussions.

Author Contributions

Conceived and designed the experiments: HC NK MB. Performed the experiments: HC MB. Analyzed the data: HC NK MB. Contributed reagents/materials/analysis tools: HC NK MB. Wrote the paper: HC MB.

References

1. Stallings RL (2007) Are chromosomal imbalances important in cancer? Trends in genetics : TIG 23: 278–283.
2. Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, et al. (2006) DNA copy number amplification profiling of human neoplasms. Oncogene 25: 7324–7332.
3. Weir BA, Woo MS, Getz G, Perner S, Ding L, et al. (2007) Characterizing the cancer genome in lung adenocarcinoma. Nature 450: 893–898.
4. Wiedemeyer R, Brennan C, Heffernan TP, Xiao Y, Mahoney J, et al. (2008) Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. Cancer cell 13: 355–364.
5. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, et al. (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature 446: 758–764.
6. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068.
7. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466: 869–873.
8. Bergamaschi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, et al. (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and geneexpression subtypes of breast cancer. Genes, chromosomes & cancer 45: 1033–1040.
9. Hu X, Stern HM, Ge L, O'Brien C, Haydu L, et al. (2009) Genetic alterations and oncogenic pathways associated with breast cancer subtypes. Molecular cancer research : MCR 7: 511–522.
10. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science (New York, NY) 258: 818–821.
11. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature genetics 23: 41–46.
12. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome research 14: 287–295.
13. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative metaanalysis of chromosomal CGH data. BMC cancer 7: 226.
14. Alloza E, Al-Shahrour F, Cigudosa JC, Dopazo J (2011) A large scale survey reveals that chromosomal copy-number alterations significantly affect gene modules involved in cancer initiation and progression. BMC medical genomics 4: 37.
15. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic acids research 39: D1005–10.
16. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, et al. (2010) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic acids research 39: D1002–D1004.
17. Scheinin I, Myllykangas S, Borze I, Böhling T, Knuutila S, et al. (2008) CanGEM: mining gene copy number changes in cancer. Nucleic acids research 36: D830–5.
18. Cao Q, Zhou M, Wang X, Meyer CA, Zhang Y, et al. (2011) CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic acids research 39: D968–74.
19. Baudis M, Cleary ML (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics (Oxford, England) 17: 1228–1229.
20. Baumbusch LO, Aarøe J, Johansen FE, Hicks J, Sun H, et al. (2008) Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. BMC genomics 9: 379.
21. Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, et al. (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC genomics 10: 588.
22. Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, et al. (2007) A comparison of DNA copy number profiling platforms. Cancer research 67: 10173–10180.
23. Bengtsson H, Ray A, Spellman P, Speed TP (2009) A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. Bioinformatics (Oxford, England) 25: 861–867.
24. Heinrichs S, Look T (2007) Identification of structural aberrations in cancer by SNP array analysis. Genome biology. pp 1–5.
25. Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics 39: S16–S21.
26. Lubin M, Lubin A (2009) Selective killing of tumors deficient in methylthioadenosine phosphorylase: a novel strategy. PloS one 4: e5735.
27. Krasinskas AM, Bartlett DL, Cieply K, Dacic S (2010) CDKN2A and MTAP deletions in peritoneal mesotheliomas are correlated with loss of p16 protein expression and poor survival. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 23: 531–538.
28. Smith JS, Tachibana I, Passe SM, Huntley BK, Borell TJ, et al. (2001) PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. Journal of the National Cancer Institute 93: 1246–1256.
29. Li J (1997) PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. Science (New York, NY) 275: 1943–1947.
30. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, et al. (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proceedings of the National Academy of Sciences of the United States of America 103: 17402–17407.
31. Zhang W, Zhu J, Bai J, Jiang H, Liu F, et al. (2010) Comparison of the inhibitory effects of three transcriptional variants of CDKN2A in human lung cancer cell line A549. Journal of experimental & clinical cancer research : CR 29: 74.
32. van der Rhee JI, Krijnen P, Gruis NA, de Snoo FA, Vasen HFA, et al. (2011) Clinical and histologic characteristics of malignant melanoma in families with a germline mutation in CDKN2A. Journal of the American Academy of Dermatology.
33. Bourdeaut F, Isidor B, Ferrand S, Thomas C, Moreau A, et al. (2011) Homozygous PTEN deletion in neuroblastoma arising in a child with Cowden syndrome. American journal of medical genetics Part A 155: 1763–1766.
34. Jin K, Kong X, Shah T, Penet MF, Wildes F, et al. (2011) Breast Cancer Special Feature: The HOXB7 protein renders breast cancer cells resistant to tamoxifen through activation of the EGFR pathway. Proceedings of the National Academy of Sciences of the United States of America.
35. Wiltshire RN, Rasheed BK, Friedman HS, Friedman AH, Bigner SH (2000) Comparative genetic patterns of glioblastoma multiforme: potential diagnostic tool for tumor classification. Neurooncology 2: 164–173.
36. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic acids research 39: D876–82.
37. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research 39: D945–50.
38. Gearhart J, Pashos EE, Prasad MK (2007) Pluripotency redux–advances in stem-cell research. The New England journal of medicine 357: 1469–1472.
39. Dalla-Favera R, Bregni M, Erikson J, Patterson D, Gallo RC, et al. (1982) Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. Proceedings of the National Academy of Sciences of the United States of America Vol. 79: 7824–7827.
40. Climent J, Dimitrow P, Fridlyand J, Palacios J, Siebert R, et al. (2007) Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. Cancer research 67: 818–826.
41. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, et al. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer cell 10: 529–541.
42. Stevens KN, Fredericksen Z, Vachon CM, Wang X, Margolin S, et al. (2012) 19p13.1 is a triple negative-specific breast cancer susceptibility locus. Cancer research.
43. Park NI, Rogan PK, Tarnowski HE, Knoll JHM (2012) Structural and genic characterization of stable genomic regions in breast cancer: Relevance to chemotherapy. Molecular oncology.
44. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, et al. (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics 84: 524–533.

45. Bengtsson H, Simpson K, Bullard J, Hansen K (2008) aroma.affymetrix: A genetic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Tech Report #745 Department of Statistics, University of California, Berkeley.
46. Bengtsson H, Wirapati P, Speed TP (2009) A single-array preprocessing method for estimating fullresolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics (Oxford, England) 25: 2149–2156.
47. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, et al. (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology 34: 591–602.
48. F C, AL A, SA K, TP S, VL SM (2005) NUSE and RLE: Quality assessment of oligonucleotide microarray data to quantify systemic variation. 2005 Meeting of the Federation of Clinical Immunology Societies Boston, MA.

12.2 Publication 5: Reverse Phase Protein Arrays identify mechanisms of PKC- dependent radio-resistance in primary human fibroblasts (Manuscript in preparation)

Reverse Phase Protein Arrays identify mechanisms of PKC-dependent radio-resistance in primary human fibroblasts

Andrej Bluwstein[1,2], Nitin Kumar[3], Karolin Meyer[1,2], Jens Traenkle[4], Jan van Oostrum[4], Hubert Rehrauer[5], Michael Baudis[3] and Michael O. Hottiger*

Running title: Proteomics of irradiation induced DNA damage

[1]Institute of Veterinary Biochemistry and Molecular Biology (IVBMB), University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[2]Cancer Biology PhD Program, Life Science Zurich Graduate School, University of Zurich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[3]Institute for Molecular Life Science (IMLS),

[4]Bayer Technology Services GmbH, Zeptosens Platform, Leverkusen, Germany

[5]Functional Genomics Center Zurich (FGCZ),

*Corresponding author: Michael O. Hottiger, hottiger@vetbio.uzh.ch

Key words: DNA damage, irradiation, PKC, sensitization, reverse phase protein array

Running title: IR induces pro-survival signaling through the PKC pathway

Abstract

The DNA damage response has been mainly studied by investigating the effect of genotoxic agents such as ionizing radiation on the structure and function of chromatin. In the global approach described here, reverse phase protein arrays coupled with kinetic analyses were employed to decipher the cytoplasmic signaling network that is induced upon genotoxic stress by ionizing radiation (IR).

Proteome changes that have not been implicated in cellular IR-induced response before included the de-phosphorylation of Akt and c-Myc, a decrease of Rb, the activation of PAK4/5/6 as well as the early up-regulation of CREB, Bcl-2 and the delayed activation of PKC-zeta/lambda, Bad and Bcl-xl. Importantly, PKC signaling acted upstream of ATM, the transcription factor CREB, ERK, Bad and Bcl-2 and thereby regulated pro-survival and apoptotic pathways. In agreement with this finding, inhibition of PKC enzymatic activity hypersensitized MRC-5 cells to IR treatment by inducing CREB and Bad dephosphorylation, by increasing p53 stabilization, Bad up-regulation and Bcl-2 down-regulation. Together, our analysis strengthens the importance of cytoplasmic signaling induced by IR and suggests a key role of PKC signaling in the conditioning of stressed cells to allow efficient DNA repair and to prevent cell death and apoptosis.

Introduction

Genotoxic stress such as ionizing radiation (IR) alters the organization and function of the genetic material and thereby induces diverse cellular responses that culminate in the activation of cell-cycle checkpoints or DNA damage response (DDR) pathways (1-3). A hallmark of DDR signaling is the slowdown or arrest of the cell-cycle progression by inhibiting entry into S-phase and mitosis through the activation of cell-cycle checkpoint signaling cascade (at the G1/S or G2/M phase checkpoints) (4, 5), which is regulated by a complex signaling network including the established DDR pathways ATM and ATR (6). Depending on the type of DNA lesion, different repair mechanisms are activated and a defect in any of these can cause severe syndromes such as tissue degeneration, sensitivity to DNA-damaging agents and predisposition to cancer (7, 8).

Recent large-scale mass spectrometry based analyses have significantly extended the understanding of DDR signaling and identified hundreds of new phosphorylation sites and target proteins (9-11). Zeptosens reverse phase protein arrays (RPPA) represent a new and powerful technology that relies on planar waveguide nanostructured chips for high sensitivity detection and the high-throughput quantification of protein changes at the single cell level in a multiplex setting (12). This approach allows performing quantitative kinetic analyses of hundreds of proteins in the same cell or tissue extract and thus generates data equivalent of thousands of western blots within a week (13). Protein arrays have already been applied to study changes of expression level or posttranslational modification status of proteins in cellular stress conditions or in normal and cancerous prostate and ovarian tissue in smaller scale (14-16).

Here, we applied the RPPA technology to study IR induced stress pathways in primary cells on a "signalosome level" by analyzing cellular candidate markers. In contrast to earlier studies, this work was performed with primary MRC-5 human fibroblasts and 165 different antibodies were

included in order to quantify and study all major components and post-translational modifications of nuclear and cytoplasmic stress signaling events. These results implicated significantly changed kinetics of pro-survival and anti-apoptotic key determinants in cell cycle arrest and the ability to withstand severe DNA damage.

Besides known components of the DNA damage response, such as changes in p21, p38 and p53 well as activation of the MEK1/2, ERK1/2 and LKB-AMPK pathways, the RPPA analysis described here also identified novel components of the cellular IR response. Among these changes were the de-phosphorylation of Akt and c-Myc, a decrease of Rb, the activation of PAK4/5/6 as well as the up-regulation of CREB, Bad, Bcl-2, Bcl-xl and PKC-zeta/lambda. The PKC signaling pathway is known to integrate extracellular signals, calcium and secondary lipid messengers to regulate diverse cellular responses including apoptosis during genotoxic stress (17, 18). The physiological importance of PKC enzymatic activity was highlighted by the hypersensitive phenotype of PKC inhibitor treated MRC-5 cells, which has been reported earlier (19). However, the mechanism by which PKC signaling leads to an IR-hyperresistant phenotype in primary human fibroblasts was so far unknown. Here, we elucidated the PKC-dependent pro-survival and anti-apoptotic signals that confer the high IR-resistance to MRC-5 fibroblasts. PKC acted upstream of ATM, the cAMP response element-binding protein (CREB) transcription factor, ERK and Bad and Bcl-2 and thereby, regulated pro-survival and apoptotic pathways. Consequently, PKC inhibition hypersensitzed MRC-5 cells to IR and led to the induction of pro-apoptotic factors (Bad) and the down-regulation of anti-apoptotic markers (Bcl-2 and p53 stabilization). This multiplex analysis of the IR-induced signaling network thus identified PKC as a new key denominator of the IR response that conditions stressed cells for efficient DNA repair and thereby prevents cell death and apoptosis.

Results

Ionizing radiation induces significant proteome changes in MRC-5 fibroblasts

Kinetic quantitative proteome changes were analyzed with reverse phase protein arrays (RPPA) after ionizing radiation (IR, 10 and 40 Gy) exposure of MRC-5 fibroblasts (Figs. 1A, S1). Irradiation of MRC-5 cells with IR doses between 4 - 10 Gy induced a dose-dependent increase and time-dependent decrease of H2A.X phosphorylation at Ser139 (γH2A.X) (Figs. S2A-C). However, DNA strand breaks were efficiently repaired within 8 h post-irradiation and continued 24-48 h following IR treatment, as indicated by the strong reduction in γH2AX and 53BP1 foci (Fig. S2C). Even after 40 Gy irradiation, cells did not exhibit elevated apoptosis as indicated by the stable fraction of Annexin V positive cells and undetectable expression of NOXA and PUMA (relative to a significant increase after staurosporine or cycloheximide treatment; Figs. S3A-D). These results documented the efficient DDR of human MRC-5 lung fibroblasts (20) and confirmed the suitability of 10 Gy and 40 Gy IR to induce a DDR in MRC-5 cells. Based on these data it was concluded that 10 Gy and 40 Gy IR induce specific, significant and reproducible non-lethal changes in the proteome of MRC-5 fibroblasts.

The arrays were probed with 165 validated antibodies that specifically recognize proteins of canonical signaling pathways. All quantitative RPPA data were subjected to statistical analysis (ANOVA: $p < 0.05$, $n = 3$ and fold change cut off: fold change $\geq$ 1.5xSD) (Figs. S4A-B). Altogether, 78 protein changes were either significant based on ANOVA and/or passed the 1.5xSD cut off (Figs. S5-S8).

Ionizing radiation of MRC-5 cells activates classical DNA damage response pathways

To validate the RPPA analysis of IR-treated MRC-5 fibroblasts, known components of the DDR pathway were assessed. Upon 10 Gy IR, levels of phosphorylated p38 MAPK (pT180, pY182), the effector Hsp27 (pS78), p53 (pS15) and Chk1 (pS345), as well as total p53 and p21 amounts, showed similar profiles (clusters 1-3, Figs. 1B, S9A). Activation of the cytoplasmic stress kinase p38 MAPK and its downstream effector, the small heat shock protein Hsp27, were both ATM-dependent as confirmed by Western blotting (Fig. S9B). Changes in p53, which is the main target for ATM in response to IR (21), correlated with delayed up-regulation of p21.

Upon 40 Gy IR, early proteome changes comprised the phosphorylation of MEK1/2 (pS217/222, pS221/226), ERK1/2 (pT202/185, pY204/187), LKB (pS428) and cyclin D1 (pT286), which were sharply induced at 0.5 h after irradiation and relaxed to much lower and constant levels thereafter (cluster 1, Figs. 1C, S9C). Both, the MEK-ERK and the LKB-AMPK pathways signal from the cell membrane and the cytoplasm to the nucleus and have already been linked to DDR (22, 23).

The classical DDR protein changes comprising increased phosphorylation of Chk1 (pS345) and p53 (pS15), as well as higher protein levels of p53 and its downstream target p21 (CIP/WAF1), were only observed at later time-points after 40 Gy IR (clusters 2-4). Phosphorylation of p38 MAPK (pT180, pY182) and its phosphorylation target MAPKAP2 (pT334) showed even an early decline, followed by a transient increase after 40 Gy IR (cluster 5). Similarly, HSP27 and Chk1 were phosphorylated, while Chk2 phosphorylation upon 40Gy IR could only be shown by Western Blotting due to poor antibody performance in the RPPA (Fig. S9D). These results indicated that the activation of classical DDR components is preceded and likely regulated by cytoplasmic signaling cascades.

The essential role of the p53 pathway in DDR and its function as a "guardian of the genome" (24, 25) is emphasized by the fact that it was the only pathway significantly overrepresented among the proteins responding to both, 10 Gy and 40 Gy IR (Fig. 2A) and its many interactions (more than 14 interactions) in the predicted protein network (Figs. 2B, S9E). Protein network analysis also predicted a highly cross-linked network of IR-induced protein changes (e. g. Bad, SRC, ERK (MAPK1, MAPK3) that function in cytoplasmic signaling pathways.

In conclusion, these findings supported the validity of the protein array approach described here and implicated cytoplasmic pro-survival signaling in the DDR upon IR treatment.

10 Gy and 40 Gy IR activate novel nuclear regulators in MRC-5 cells

DDR can either slow-down cell cycle progression to allow efficient DNA repair or induce apoptosis. However, IR treatment (10 and 40 Gy) of MRC-5 cells neither induced apoptosis (Fig. S2A-C) nor a G1/S cell cycle arrest (Fig. S9F), even though p53 and p21 up-regulation and Cyclin D1 phosphorylation (pT286) and degradation (Figs. 1B-C) hinted at such a possibility. Therefore, novel IR-induced protein changes functioning in the cell cycle were evaluated.

Novel IR-dependent changes in the nucleus included a significant down-regulation of pRb (Rb) protein levels (cluster 5, Figs. 1B, S9A), which may explain the absence of a G1/S cell cycle arrest (26). Reduction in Rb could be confirmed by Western blotting and quantitative RT-PCR analysis (Figs. 3A, S9D, S9G), which supports *in vivo* Rb down-regulation upon IR in MRC-5 fibroblasts. The relative phosphorylation of Rb at S780 increased significantly increased upon 10 Gy after 8 h and already after 2 h following irradiation with 40 Gy, which could explained why a G1/S cell cycle arrest was not observed (Fig. S9F). Other novel nuclear and IR-dependent

changes included the increased expression of Cyclin C and the translation initiation factor elF4E, which are both involved in cell cycle regulation.

Instead of a G1/S cell cycle arrest, 10 Gy and 40 Gy IR led to a visible G2/M cell cycle arrest after already 24 h post-IR. Interestingly, a strong reduction in colony formation was observed even after low doses of IR (1-6 Gy), indicating cell cycle exit (most probably from the G2/M cell cycle phase) and cell senescence after extended periods of time (Fig. S9H). Despite the efficient repair of DNA double strand breaks induced by IR within 8h (Fig. S2B), residual breaks that can potentially cause a permanent cell cycle arrest or replicative senescence cannot be excluded. For instance, residual γH2AX foci were detected even after irradiation with 1 Gy (Fig. S2A), which explains the observed reduction in colony formation under the tested conditions.

These findings hint at a mechanism, which allows primary cells to survive severe DNA damage, while senescence prevents transmission of mutations into the next generation by partly escaping from the cell cycle.

Novel IR-induced protein changes are involved cytoplasmic pro-survival signaling

The RPPA data for known DDR components indicated that cytoplasmic factors responded more quickly than nuclear signaling proteins to IR treatment (Figs. 1B-C, S9A-C). In addition, cytoplasmic signaling has already been described to regulate the DDR and DNA repair in a break independent manner (27, 28). Therefore, the novel IR-dependent cytoplasmic signaling routes were analyzed in more detail.

Several novel components of cytoplasmic stress signaling involved in the regulation of cell survival were up-regulated at both 10 Gy and 40 Gy IR. These changes included early phosphorylation of CREB (pS133), Bad (pS112 and pS136), Bcl-2 (pS70 for 10Gy only) and a

cluster comprising protein kinase C (PKC) family members with phosphorylation of the novel PKC-δ (pT505) and of the atypical PKC-ζ/λ (pT410/403) in the activation loop, which marks the activated state of these two protein kinase C isoforms (Figs. 1B, 1C, 3B) (29). Among the late-responding protein changes upon 40 Gy was the phosphorylation of PAK4/5/6 (p21-activated kinases of the group II) (cluster 4, Figs. 1C, S9C), which are kinases playing a role in cytoskeletal reorganization and survival that were not linked to DDR previously (30). Western blotting also showed a late increase in STAT3 (pS727) phosphorylation, which is an anti-apoptotic modification described to be downstream of growth factor signaling and so far not implicated in the DDR (31).

Together, these analyses suggested that IR and the consecutive DNA damage broadly affect cytoplasmic pro-survival signaling network. The majority of the identified signaling pathways mediated by MEK-ERK, p38-MK2-Hsp27 and PKCs function as cytoplasmic pro-survival pathways, indicating that IR treatment of MRC-5 cells induces cell protective regulatory networks.

PKC inhibition sensitizes MRC-5 cells to IR

Based on our initial findings and to elucidate the mechanism leading to elevated IR resistance of primary human fibroblasts, we performed a radiosensitivity screen using inhibitors specific for key components of the p38, ERK, JNK, AMPK and PKC pathways (Figs. S10A-B). Inhibition of p38 and MEK did not reveal a significant reduction in cell viability upon IR, while JNK and AMPK inhibitors strongly reduced MRC5 survival, but independent of IR. In contrast, MRC-5 cells that were pretreated with PKC inhibitors (GF109203X or RO-318220) showed a stronger reduction in cell viability than cells pretreated with DMSO (the solvent) alone or without IR

exposure (Figs. 4A, S10C-D, S11A). These results suggested that inhibition of PKC signaling in MRC-5 cells leads to radiosensitivity. To elucidate whether PKC signaling acts upstream of the DDR in IR-treated MRC-5 cells, ATM phosphorylation as well as p53, p21 and γH2A.X were assessed (Figs. 4B, S11B-C). PKC pan inhibition by GF109203X resulted in reduced ATM phosphorylation, reduced p53 and p21 levels and slight reduction in γH2A.X phosphorylation after 2-4 h recovery from IR. A similar effect could be observed using Ro-318220, an alternative PKC-pan inhibitor, which also sensitized MRC-5 cells to 10 and 40 Gy (Figs. S10C-D). These results placed PKC signaling upstream of the ATM pathway following IR treatment, indicating induction of the DDR by signals from the cytoplasm.

In order to decipher the signaling cascade leading to MRC-5 hypersensitization upon PKC inhibitor treatment, phosphorylation of the pro-apoptotic factor Bad and the transcription factor CREB was analyzed (Fig. S10C). PKC inhibition caused reduced phosphorylation of CREB (pS133) and Bad (pS136), which directs cells towards apoptosis.

To confirm apoptosis as the mechanism of radiosensitization by PKC inhibition, changes in pro- and anti-apoptotic markers were assessed by immunoblotting upon combination treatment (Fig. 4C). 40 Gy IR treatment in the presence of PKC inhibitors led to markedly up-regulated p53 levels, which was not observed in response to 10 Gy IR (Fig. 4C). The pro-apoptotic marker Bad was significantly induced in the presence of PKC inhibitors and IR treatment, while the anti-apoptotic marker Bcl-2 was drastically reduced. Strikingly, 40 Gy induced elevated expression of the anti-apoptotic factor Bcl-xl following 20 h recovery, confirming RPPA data which revealed late Bcl-xl up-regulation only upon 40 Gy (Figs. S6, S8, S11D). Even more interesting, PKC inhibition led to a significant reduction in Bcl-xl, indicating PKC-dependent survival signaling especially upon severe DNA damage (Fig. 11D). Activation of an apoptotic program was also

confirmed by increased ARTD1 (PARP-1) cleavage as indicated by a reduction in ARTD1 levels and accumulation of the 89 kDa cleavage fragment specifically upon IR in combination with PKC inhibitor treatment (Fig. 4C-D, S11D).

In summary, these results suggest that DDR upon IR treatment involves activation of cytoplasmic PKC signaling, which acts upstream of the ATM pathway and the consecutive activation of pro-survival signaling events (MEK/ERK, CREB, Bcl-2 and Bcl-xl) as well as inactivation of pro-apoptotic factors (Bad) (Fig. 5). Activation of cytoplasmic PKC signaling upon IR is thus a novel mechanism that orchestrates the different IR-induced responses to ensure cell survival of primary human fibroblasts.

Discussion

To understand the dynamics of *in vivo* signaling events in response to IR, we performed a kinetic RPPA analysis, which is an antibody-based, targeted, proteomic approach allowing quantification of proteome changes at the cellular level in a multiplex setting (32). In contrast to earlier similar approaches, this study included many cytoplasmic signaling components and was thus not limited to nuclear proteins. The antibodies used here were specific for nuclear and cytoplasmic proteins, whole protein levels as well as different protein modifications and thereby elucidated the extended signaling network that is activated upon irradiation. We used MRC-5 primary human lung fibroblasts as a model cell line, because these cells have been previously employed for studying DNA repair after the induction of double strand breaks by IR and are known to be less sensitive to IR than transformed cells (20). Although primary MRC-5 fibroblasts can efficiently repair DNA damage induced by IR doses up to 80 Gy (20), it was so far not clear how these cells prevent apoptosis, which is induced at much lower IR doses in other cell types (33).

One important advantage of reverse phase protein arrays over other proteomic approaches is the study of many treatment conditions in parallel. We were able to describe a dynamic picture of IR-dependent proteome changes upon 10 Gy and 40 Gy in a multiplex setting. Strikingly, IR induced an early response, which was governed predominantly by growth factor and cytokine dependent signaling pathways including members of the MAPK family (MEK/ERK), protein kinase C (PKC) family (PKCδ, PKC-ζ/λ, PKCβII) as well as anti- and pro-apoptotic Bcl-2 family members (Bcl-2, Bad). Except for early Chk1 and Cyclin D1 phosphorylation, major ATM-dependent signaling events (H2AX, p53, p21, MKK3/6, p38, MK2, Hsp27, Chk2) followed delayed kinetics peaking 2-8 h upon IR, thus indicating cytoplasmic signaling events upstream of the canonical DDR. Indeed, an increasing body of evidence points at a cytoplasmic signaling network induced

upon IR in regulating DNA break repair (34).

The most important finding of this systematic RPPA analysis was the identification of PKC as a key regulator that orchestrates the downstream signaling pathways regulating cell cycle arrest and DNA repair, apoptosis and cellular survival (Fig. 5). IR-dependent PKC activation regulated ATM, Bad and CREB to induce DNA repair, prevent apoptosis and induce pro-survival signaling. Most proteome changes, in particular upon 40 Gy IR, indicated changes in mitogenic signaling events, which control cell-survival (MKK3/6-p38-MK2 pathway, MEK-ERK pathway, LKB-AMPK pathway, PKC-Bad pathway). The PKC signaling pathway thus acted as an upstream regulator of the DDR and as a key determinant for a cells ability to withstand and recover from IR. PKC inhibitor application prior to IR treatment stabilized p53, which is a hallmark of cell cycle arrest or apoptosis (24), induced pro-apoptotic components (ARTD1 cleavage, Bad) and down-regulated anti-apoptotic factors (Bcl-2 and Bcl-xl). Our findings thus suggest that hypersensitization of primary fibroblasts by PKC inhibitors induces an apoptotic program that is mediated by effects on ATM, CREB, Bad and pro-apoptotic genes, particularly upon 40Gy. This is in line with cell viability data showing a stronger decline in cell viability upon 40Gy in combination with PKC inhibition compared to the lower dose. PKC-dependent regulation of DNA repair through the ATM pathway is thus proposed as a part of the mechanism by which primary fibroblasts resist high IR doses and circumvent apoptosis, which is in agreement with the high DNA repair efficiency and the PKC-dependent regulation of p53 function (20, 35).

This result has important implications for our understanding of DDR and the regulation of nuclear events in general and supports the concept of the cellular, in contrast to the nuclear, radiation response (36). Cellular stress such as IR is perceived at the plasma membrane and in the cytoplasm and initiates cytoplasmic signaling pathways that consecutively control nuclear events

such as the cell cycle and DNA repair. This is also suggested by the induction of IR-induced signaling in cells exposed to IR conditioned medium, which is termed the "bystander effect" (37). The classical DDR, comprised of nuclear events regulating cell cycle progression and DNA repair, is thus preeceded and controled by the cytoplasmic radiation response, which is mediated at least in part by the PKC signaling pathway.

This RPPA analysis of the IR-induced signaling network identified a new, cytoplasmic regulation of the DDR response that conditions stressed cells for efficient DNA repair and thereby prevents cell death and apoptosis. These findings may explain how tumorigenesis is initiated in response to severe irradiation stress; namely by the induction of pro-survival signaling and the evasion of apoptosis, which gives cells time to accumulate the mutations eventually leading to cancer. The results presented here thus shed new light on the cellular machinery that enables cells to withstand IR and may thus help optimize and improve radiotherapy.

Materials and Methods

Cell culture, treatment, lysis and viability assays

MRC-5 and WI-38 human lung fibroblast cell lines (38, 39) were obtained from the American Type Culture Collection (ATCC) and cultured in supplemented MEM (Invitrogen). Cells were exposed to ionizing radiation (IR) using an X-ray generator (Pantak Seifert X-ray System; 120 kV; 19 mA; aluminium filter, 3.11 Gy/min) and recovered for different time-periods. Whole cell extracts were prepared with Zeptosens Cell Lysis Buffer CLB1 (Bayer Technology Services GmbH) or RIPA lysis buffer. Cell viability was determined by seeding $2x10^3$ cells in 96 well plates over night before irradiation or drug treatment. Cell viability was quantified using the WST-1 proliferation reagent (Roche) and a plate reader (Tecan). The clonogenic assay was performed as described elsewhere (40).

Reverse Phase Protein Arrays

RPPA were prepared as described (13) (32) (see supplementary material). The eight data points (100, 75, 50, 25% lysate amount in duplicates) were fitted using a weighted linear least squares fit (41) and the relative fluorescence intensity determined by interpolating at the median protein concentration or modification. To correct for small variations in protein content, relative intensities were normalized to the signals of β-Catenin, which did not show any significant variation in response to IR over indicated time points. Significance and clustering analysis was performed as described in detail in the supplementary material.

Immunoblotting

For Western Blot analysis, bands were visualized by either using horseradish peroxidase-conjugated antibodies (1:5000, GE Healthcare) and ECL detection (GE Healthcare) or by using IR-Dye-conjugated antibodies (1:15000, LI-COR) and detection by the Odyssey infrared imaging system (LI-COR). Antibodies used for Western blotting are listed in the supplementary material.

Immunofluorescence microscopy

MRC-5 cells grown on cover slips over night (≈ 1x 10^4 cells) were irradiated and immunohistochemically stained with primary (1:500 mouse anti-Histone H2A.X Phospho (Ser139) IgG1, Millipore or 1:500 rabbit anti-53BP1 IgG1, Santa Cruz) and secondary antibodies (1:250 FITC conjugated donkey anti-mouse IgG, Jackson ImmunoResearch).

Flow cytometry

FACS analysis of IR-treated or untreated MRC-5 cells was performed using a Dako CyAn ADP flow cytometer (Dako). Annexin V staining was performed using the FITC Annexin V apoptosis detection kit (BD Pharmingen).

RNA extraction and real-time PCR analysis

Total RNA was reverse transcribed using High-Capacity cDNA Reverse Transcription kit (Applied Biosystems). Real-time PCR was performed using SYBR green premixed buffer and analysed by the Rotor-Gene Q cycler (QIAGEN).

Author Contributions

AB, NK, JT and MOH designed the experiments; AB, NK and KM performed and analysed the experiments; JvO, HR, MB and MOH supervised the study; all authors contributed to the preparation of the manuscript.

Except for JT, who is an employee of Bayer Technology Services GmbH, the authors declare that they have no conflict of interest.

Acknowledgements

F. Freimoser (University of Zurich) provided editorial assistance and critical input during the writing. This work was supported in part by the Kanton of Zurich (to M.O.H.), Oncosuisse (KLS 02396-02-2009) and the UBS foundation.

References

1. Luijsterburg MS & van Attikum H (2011) Chromatin and the DNA damage response: The cancer connection. *Mol Oncol* 5(4):349-367.
2. Roos WP & Kaina B (2006) DNA damage-induced cell death by apoptosis. *Trends in molecular medicine* 12(9):440-450.
3. Harper JW & Elledge SJ (2007) The DNA damage response: ten years after. *Mol Cell* 28(5):739-745.
4. Bartek J & Lukas J (2001) Mammalian G1- and S-phase checkpoints in response to DNA damage. *Current opinion in cell biology* 13(6):738-747.
5. Lobrich M & Jeggo PA (2007) The impact of a negligent G2/M checkpoint on genomic instability and cancer induction. *Nature reviews. Cancer* 7(11):861-869.
6. Seviour EG & Lin SY (2010) The DNA damage response: Balancing the scale between cancer and ageing. *Aging* 2(12):900-907.
7. Hoeijmakers JH (2001) Genome maintenance mechanisms for preventing cancer. *Nature* 411(6835):366-374.
8. Moses RE (2001) DNA damage processing defects and disease. *Annu Rev Genomics Hum Genet* 2:41-68.
9. Bennetzen M*, et al.* (2010) Site-specific phosphorylation dynamics of the nuclear proteome during the DNA damage response. *Mol Cell Proteomics* 9(6):1314-1323.
10. Bensimon A*, et al.* (2010) ATM-dependent and -independent dynamics of the nuclear phosphoproteome after DNA damage. *Science signaling* 3(151):rs3.
11. Beli P*, et al.* (2012) Proteomic Investigations Reveal a Role for RNA Processing Factor THRAP3 in the DNA Damage Response. *Mol Cell*.

12. van Oostrum J & Voshol H (2008) Antibody-based proteomics to study cellular signalling networks. *European Pharmaceutical Review* (2):31-35.

13. Voshol H, Ehrat M, Traenkle J, Bertrand E, & van Oostrum J (2009) Antibody-based proteomics: analysis of signaling networks using reverse protein arrays. *The FEBS journal* 276(23):6871-6879.

14. Nishizuka S*, et al.* (2008) Quantitative protein network monitoring in response to DNA damage. *J Proteome Res* 7(2):803-808.

15. Paweletz CP*, et al.* (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* 20(16):1981-1989.

16. Hudson ME, Pozdnyakova I, Haines K, Mor G, & Snyder M (2007) Identification of differentially expressed proteins in ovarian cancer using high-density protein microarrays. *Proc Natl Acad Sci USA* 104(44):17494-17499.

17. Yoshida K (2011) Role for PKCδ on Apoptosis in the DNA Damage Response. *Selected Tpics in DNA Repair*, ed Chen CC (InTech, San Diego, USA), pp 293-304.

18. Newton AC (2010) Protein kinase C: poised to signal. *American journal of physiology. Endocrinology and metabolism* 298(3):E395-402.

19. Rocha S*, et al.* (2000) Protein kinase C inhibitor and irradiation-induced apoptosis: relevance of the cytochrome c-mediated caspase-9 death pathway. *Cell Growth Differ* 11(9):491-499.

20. Kuhne M*, et al.* (2004) A double-strand break repair defect in ATM-deficient cells contributes to radiosensitivity. *Cancer Research* 64(2):500-508.

21. Shiloh Y (2003) ATM and related protein kinases: safeguarding genome integrity. *Nature reviews. Cancer* 3(3):155-168.

22. Li Z*, et al.* (2006) Src tyrosine kinase inhibitor PP2 suppresses ERK1/2 activation and epidermal growth factor receptor transactivation by X-irradiation. *Biochem Biophys Res Commun* 341(2):363-368.

23. Sanli T*, et al.* (2010) Ionizing radiation activates AMP-activated kinase (AMPK): a target for radiosensitization of human cancer cells. *Int J Radiat Oncol Biol Phys* 78(1):221-229.

24. Meek DW (2009) Tumour suppression by p53: a role for the DNA damage response? *Nature reviews. Cancer* 9(10):714-723.

25. Lavin MF & Gueven N (2006) The complexity of p53 stabilization and activation. *Cell death and differentiation* 13(6):941-950.

26. Weinberg RA (1995) The retinoblastoma protein and cell cycle control. *Cell* 81(3):323-330.

27. Dittmann K, Mayer C, & Rodemann HP (2005) Inhibition of radiation-induced EGFR nuclear import by C225 (Cetuximab) suppresses DNA-PK activity. *Radiother Oncol* 76(2):157-161.

28. Golding SE*, et al.* (2007) Extracellular signal-related kinase positively regulates ataxia telangiectasia mutated, homologous recombination repair, and the DNA damage response. *Cancer Res* 67(3):1046-1053.

29. Parekh DB, Ziegler W, & Parker PJ (2000) Multiple pathways control protein kinase C phosphorylation. *EMBO J* 19(4):496-503.

30. Wells CM & Jones GE (2010) The emerging importance of group II PAKs. *Biochem J* 425(3):465-473.

31. Catlett-Falcone R*, et al.* (1999) Constitutive activation of Stat3 signaling confers resistance to apoptosis in human U266 myeloma cells. *Immunity* 10(1):105-115.

32. Pawlak M*, et al.* (2002) Zeptosens' protein microarrays: a novel high performance microarray platform for low abundance protein analysis. *Proteomics* 2(4):383-393.

33. Torudd J*, et al.* (2005) Dose-response for radiation-induced apoptosis, residual 53BP1 foci and DNA-loop relaxation in human lymphocytes. *Int J Radiat Biol* 81(2):125-138.

34. Meyn RE, Munshi A, Haymach JV, Milas L, & Ang KK (2009) Receptor signaling as a regulatory mechanism of DNA repair. *Radiother Oncol* 92(3):316-322.

35. Yoshida K (2010) Protein kinase C, p53, and DNA damage. *Protein Kinase C in Cancer Signaling and Therapy,* Current Cancer Research, ed Kazanietz MG (Springer), pp 253-265.

36. Schmidt-Ullrich RK, Dent P, Grant S, Mikkelsen RB, & Valerie K (2000) Signal transduction and cellular radiation responses. *Radiat Res* 153(3):245-257.

37. Baskar R, Balajee AS, Geard CR, & Hande MP (2008) Isoform-specific activation of protein kinase c in irradiated human fibroblasts and their bystander cells. *The international journal of biochemistry & cell biology* 40(1):125-134.

38. Jacobs JP, Jones CM, & Baille JP (1970) Characteristics of a human diploid cell designated MRC-5. *Nature* 227(5254):168-170.

39. Hayflick L & Moorhead PS (1961) The serial cultivation of human diploid cell strains. *Experimental cell research* 25:585-621.

40. Franken NA, Rodermond HM, Stap J, Haveman J, & van Bree C (2006) Clonogenic assay of cells in vitro. *Nature protocols* 1(5):2315-2319.

41. Bevington PR (2002) *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, NY); 3rd Ed. Ed p 352.

Figure Legends

Fig. 1. Identification of significant proteome changes in response to IR.

A. Workflow for the characterization of IR-induced proteome changes by reverse phase protein arrays (RPPA). Early passage of 80-90% confluent MRC-5 cells were left untreated (U) or irradiated in biological triplicates with 10 Gy or 40 Gy. At different time-points, cell lysates were prepared and spotted in equal amounts on hydrophobic glass slides. Each lysate was spotted in a serial dilution at 1, 0.75, 0.5 and 0.25 relative to the total protein concentration in duplicates. Each array was incubated with a different primary antibody, directed against a protein of interest, and a secondary fluorophore-labeled antibody. Relative fluorescence intensities were quantified and used for statistical data analysis to identify significant proteome changes in response to irradiation. **B-C.** Time profile clustering of IR-dependent changes (ANOVA $p < 0.05$, 1.5 S.D. cut off) using Self-Organizing Tree Algorithm (SOTA) showing protein expression and modification (e.g. phosphorylation) in green or downregulation and de-modification (e.g. dephosphorylation) in red. **B.** Clustering of proteome changes upon irradiation with 10 Gy. **C.** Clustering analysis of IR-dependent proteome changes upon 40 Gy treatment.

Fig. 2. Pathway and protein-protein interaction analysis of significant IR-induced proteome changes.

A. Enrichment of pathways from the pathway interaction database (PID) with significant proteome changes (ANOVA, $p < 0.05$ / 1.5 S.D.) induced by 10 Gy only (red), 40 Gy only (blue) and overlapping changes (green) using Fisher's exact test. (FDR corrected p-value cut off = 0.1). **B.** Protein-protein interaction analysis of significant proteome changes induced by 40 Gy using

STRING with a STRING score cut off > 0.7. Colors indicate phosphorylation (light blue), total protein changes (violet), phosphorylation and total protein changes (orange). Circles represent unique changes and squares overlapping changes. Large symbols indicate proteins that interact with more than 14 members.

Fig. 3. RPPA analysis identifies novel components of the DDR in IR-treated MRC-5 cells.

A. RPPA (black bars) and western blot (white bars) analysis of IR-dependent total and phosphorylated Rb in response to 10 Gy (n=3, ANOVA with Dunnett's post-hoc test, U (untreated) as reference group, * = $p < 0.05$, ** = $p < 0.01$, *** = $p = 0.001$). Western blot see Fig. S9G. **B.** Validation of novel IR-dependent proteome changes in response to 2 Gy, 10 Gy and 40 Gy X-rays. MRC-5 were treated with 100 nM Insulin (Ins), 50 μM Anisomycin (Anis), 100 nM PMA for 1 h, left untreated or irradiated and recovered for the indicated time points before cell lysis and immunoblotting. C. Positive regulation of the ATM pathway by IR-induced PKC signaling. MRC-5 were preincubated with 5 μM GF109203X (PKCi) for 2 h following irradiation, irradiated or left untreated as control and recovered for the indicated time points before cell lysis and immunoblotting. α-ERK was used as loading control.

Fig. 4. Inhibition of IR-dependent PKC signaling reduces proliferation and survival in primary human fibroblasts in response to X-rays.

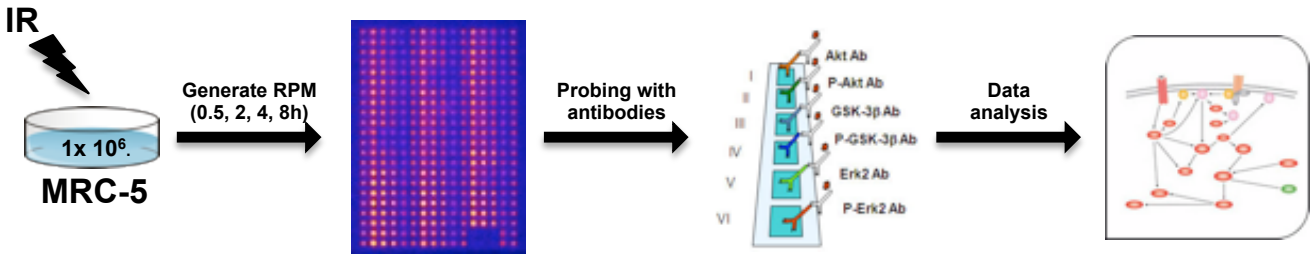
A. Radiosensitization of MRC-5 cells by a combination treatment with PKC inhibition and IR. MRC-5 were pretreated with PKCi (GF109203X, 10 μM, for 2 h) or DMSO as control before irradiation with 10 Gy or 40 Gy and recovery for the indicated time points, followed by cell viability analysis using the WST-1 assay. **B.** IR-dependent ATM pathway and pro-survival

signaling is negatively regulated by PKC inhibition. WB analysis of irradiated MRC-5 cells using ATM- and PKC-specific inhibitors (KU55933, 5 μM and Ro-318220, 5 μM with 2 h preincubation before irradiation). β-Tubulin was used as loading control. **C.** PKC inhibition and IR leads to ARTD1 cleavage, Bcl-2 down-regulation and Bad upregulation. WB analysis as in B. **D.** Quantification of ARD1 cleavage shown in C.

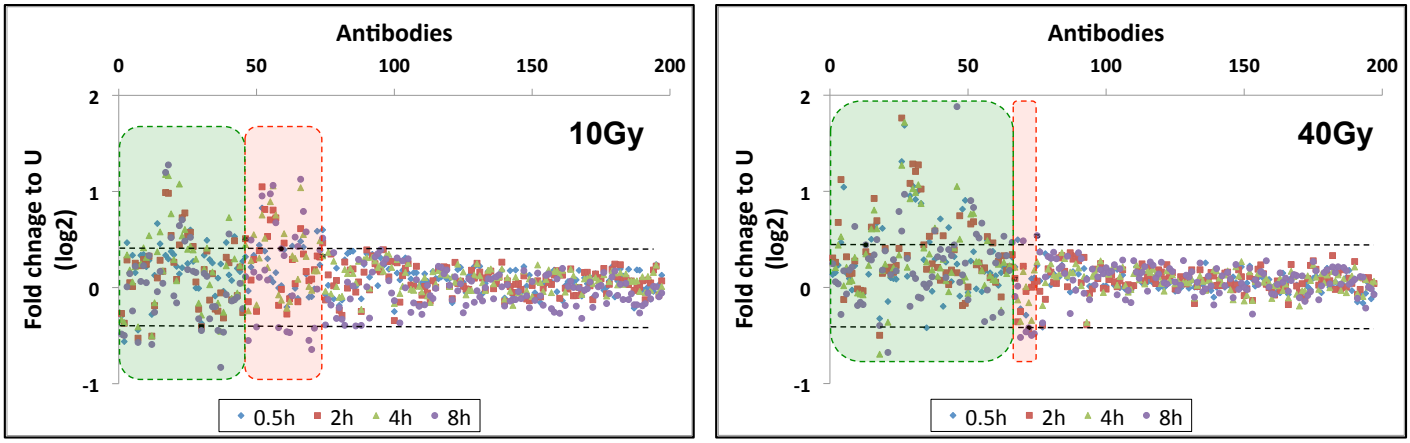
Figur 5. Model of PKC-dependent cellular pro-survival signaling in response to IR.

IR-induced PKC signaling orchestrates cell cycle arrest and DNA repair, apoptosis and cellular survival via ATM phosphorylation, p53 stabilization and p21 and γH2AX induction as well as via up-regulation of Bad, ARTD1 cleavage and activation of Bcl-2 and CREB. Hypersensitization with PKC inhibitors (GF109203X, RO-318220) leads to an IR-induced activation of apoptosis.

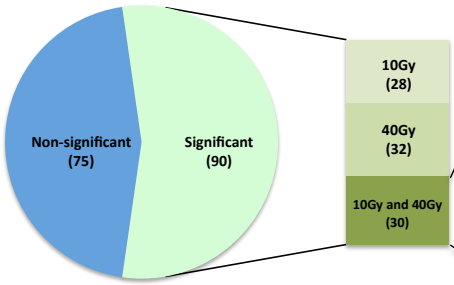
A



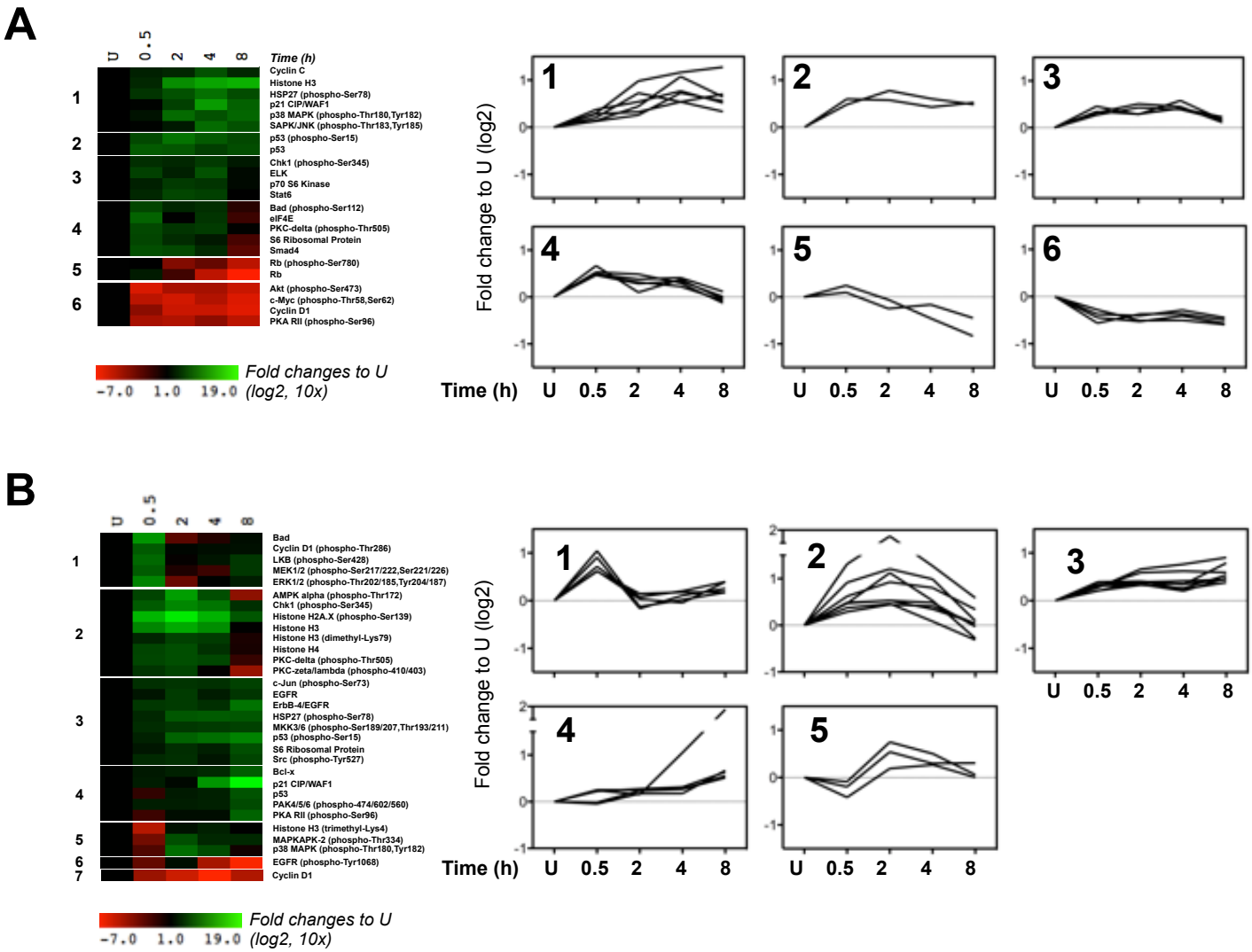
B



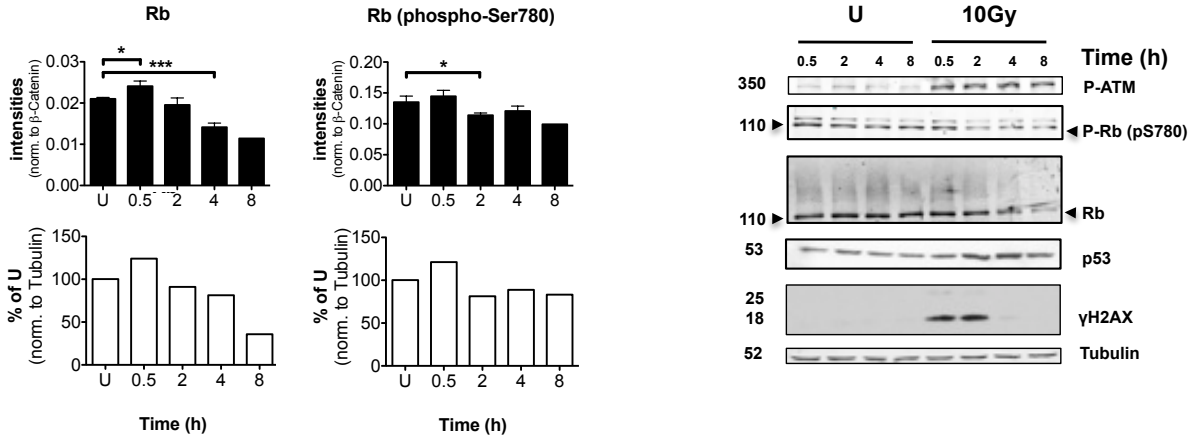
C



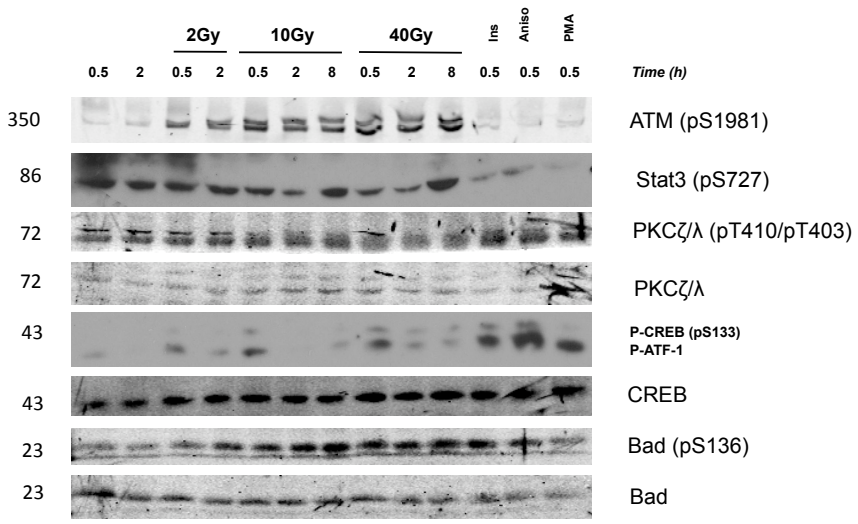
10 and 40Gy	
Canonical	**Non-canonical**
Chk1 (pS345)	Bad (pS112)
Cyclin D1	HSP27 (HSPB1) (pS78)
p21 CIP/WAF1	p38 MAPK (pT180,pTyr182)
p53	PKC-delta (pT505)
p53 (pS15)	PKA RII (p-Ser96)
…	…

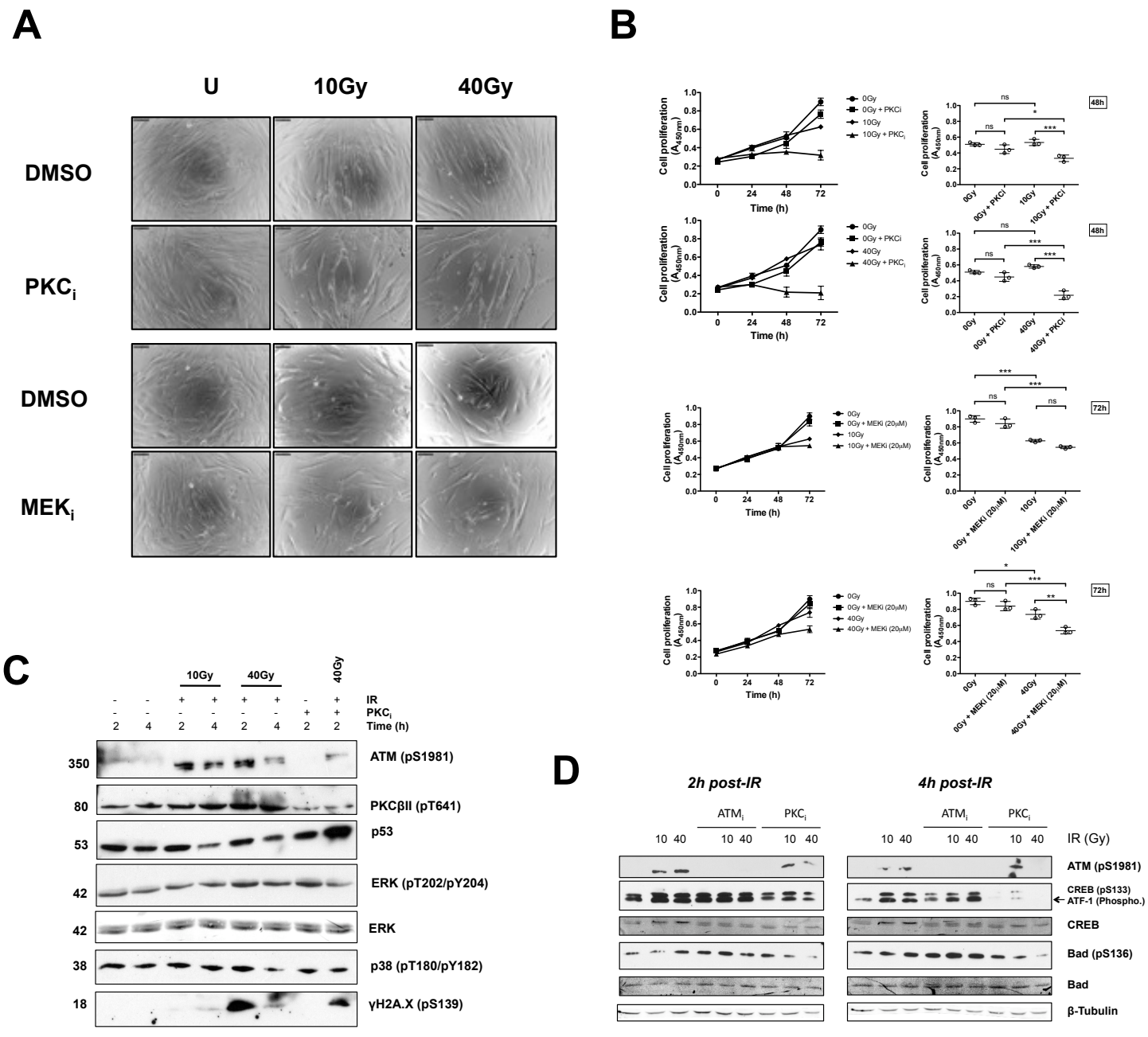

A
U
0.5
2
4
8
Time (h)
Cyclin C
Histone H3
HSP27 (phospho-Ser78)
p21 CIP/WAF1
p38 MAPK (phospho-Thr180,Tyr182)
SAPK/JNK (phospho-Thr183,Tyr185)
p53 (phospho-Ser15)
p53
Chk1 (phospho-Ser345)
ELK
p70 S6 Kinase
Stat6
Bad (phospho-Ser112)
eIF4E
PKC-delta (phospho-Thr505)
S6 Ribosomal Protein
Smad4
Rb (phospho-Ser780)
Rb
Akt (phospho-Ser473)
c-Myc (phospho-Thr58,Ser62)
Cyclin D1
PKA RII (phospho-Ser96)
1
2
3
4
5
6
Fold changes to U
-7.0 1.0 19.0 (log2, 10x)
Fold change to U (log2)
Time (h)
B
Bad
Cyclin D1 (phospho-Thr286)
LKB (phospho-Ser428)
MEK1/2 (phospho-Ser217/222,Ser221/226)
ERK1/2 (phospho-Thr202/185,Tyr204/187)
AMPK alpha (phospho-Thr172)
Chk1 (phospho-Ser345)
Histone H2A.X (phospho-Ser139)
Histone H3
Histone H3 (dimethyl-Lys79)
Histone H4
PKC-delta (phospho-Thr505)
PKC-zeta/lambda (phospho-410/403)
c-Jun (phospho-Ser73)
EGFR
ErbB-4/EGFR
HSP27 (phospho-Ser78)
MKK3/6 (phospho-Ser189/207,Thr193/211)
p53 (phospho-Ser15)
S6 Ribosomal Protein
Src (phospho-Tyr527)
Bcl-x
p21 CIP/WAF1
p53
PAK4/5/6 (phospho-474/602/560)
PKA RII (phospho-Ser96)
Histone H3 (trimethyl-Lys4)
MAPKAPK-2 (phospho-Thr334)
p38 MAPK (phospho-Thr180,Tyr182)
EGFR (phospho-Tyr1068)
Cyclin D1
1
2
3
4
5
6
7
Fold changes to U
-7.0 1.0 19.0 (log2, 10x)
Fold change to U (log2)
Time (h)
C
-log10(corrected p values)
10 Gray
40 Gray
Overlap
Angiopoietin receptor Tie2-mediated signaling
p53 pathway
Alpha-synuclein signaling
CDC42 signaling events
Cellular roles of Anthrax toxin
CXCR3-mediated signaling events
Endothelins
EPHB forward signalin
ErbB1 downstream signaling
Netrin-mediated signaling event
Neurotrophic factor-mediated Trk receptor signaling
Signaling events mediated by focal adhesion kinase
Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met
Signaling events mediated by PRL
Thromboxane A2 receptor signaling
Trk receptor signaling mediated by the MAPK pathway
D
MAP2K2
PRKCZ
PAK4
PRKCD
MAPK3
SRC
MAPK1
BAD
MAP2K1
EGFR
JUN
MAP2K3
MAPKAPK2
TP53
MAPK14
CHEK1
CCND1
MAP2K6
CDKN1A
HSPB1
STK11
PRKAA1

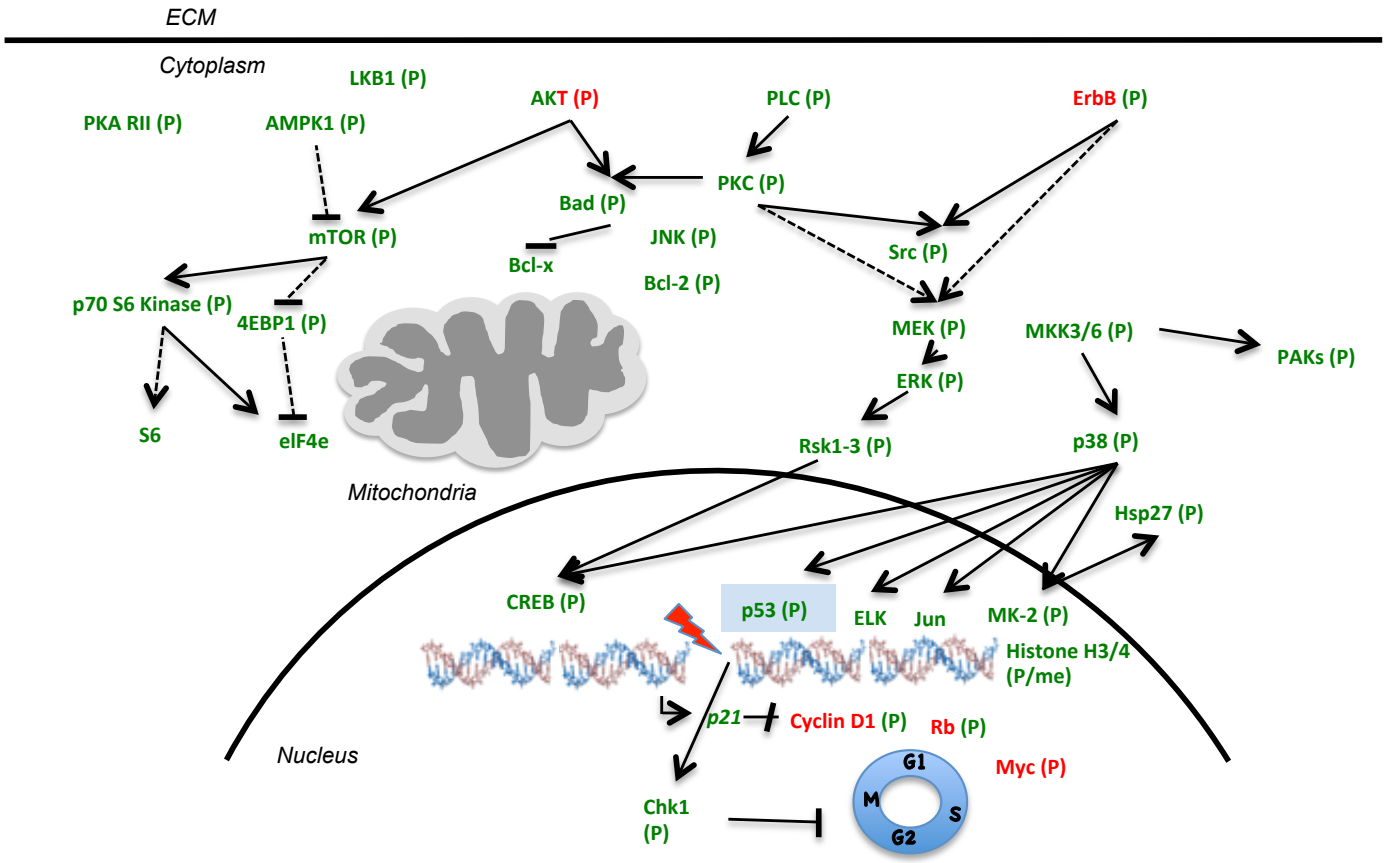
A


Rb
Rb (phospho-Ser780)
intensities (norm. to β-Catenin)
% of U (norm. to Tubulin)
Time (h)
U
10Gy
P-ATM
P-Rb (pS780)
Rb
p53
γH2AX
Tubulin

B


2Gy
10Gy
40Gy
Ins
Aniso
PMA
Time (h)
ATM (pS1981)
Stat3 (pS727)
PKCζ/λ (pT410/pT403)
PKCζ/λ
P-CREB (pS133)
P-ATF-1
CREB
Bad (pS136)
Bad


A
U
10Gy
40Gy
DMSO
PKC$_i$
DMSO
MEK$_i$
B
Cell proliferation (A_{450nm})
Time (h)
0Gy
0Gy + PKCi
10Gy
10Gy + PKC$_i$
40Gy
40Gy + PKC$_i$
0Gy + MEKi (20μM)
10Gy + MEKi (20μM)
40Gy + MEKi (20μM)
48h
72h
ns
*
**
***
C
10Gy
40Gy
40Gy
IR
PKC$_i$
Time (h)
350
80
53
42
42
38
18
ATM (pS1981)
PKCβII (pT641)
p53
ERK (pT202/pY204)
ERK
p38 (pT180/pY182)
γH2A.X (pS139)
D
2h post-IR
4h post-IR
ATM$_i$
PKC$_i$
10 40
IR (Gy)
ATM (pS1981)
CREB (pS133)
ATF-1 (Phospho.)
CREB
Bad (pS136)
Bad
β-Tubulin


ECM
Cytoplasm
LKB1 (P)
AKT (P)
PLC (P)
ErbB (P)
PKA RII (P)
AMPK1 (P)
PKC (P)
Bad (P)
mTOR (P)
JNK (P)
Bcl-x
Src (P)
Bcl-2 (P)
p70 S6 Kinase (P)
4EBP1 (P)
MEK (P)
MKK3/6 (P)
PAKs (P)
ERK (P)
S6
eIF4e
Rsk1-3 (P)
p38 (P)
Mitochondria
Hsp27 (P)
CREB (P)
p53 (P)
ELK
Jun
MK-2 (P)
Histone H3/4 (P/me)
p21
Cyclin D1 (P)
Rb (P)
Nucleus
Myc (P)
Chk1 (P)
G1
M
S
G2

13 Bibliography

[1] ACS: **Global Cancer Facts Figures 2nd Edition**. *Atlanta: American Cancer Society* 2007.

[2] Stratton MR, Campbell PJ, Futreal PA: **The cancer genome**. *Nature* 2009, **458**(7239):719–24.

[3] STM: **http://training.seer.cancer.gov/disease/categories/classification.html**. *SEER Training Modules (STM)* 20-01-2012.

[4] **http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-figures-2011** 1999.

[5] Huang JQ, Sridhar S, Chen Y, Hunt RH: **Meta-analysis of the relationship between Helicobacter pylori seropositivity and gastric cancer**. *Gastroenterology* 1998, **114**(6):1169–79.

[6] Parsonnet J, Friedman GD, Vandersteen DP, Chang Y, Vogelman JH, Orentreich N, Sibley RK: **Helicobacter pylori infection and the risk of gastric carcinoma**. *N Engl J Med* 1991, **325**(16):1127–31.

[7] Helicobacter, Group CC: **Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts**. *Gut* 2001, **49**(3):347–53.

[8] zur Hausen H: **Viruses in human cancers**. *Science* 1991, **254**(5035):1167–73.

[9] of STD Prevention D: **Prevention of genital HPV infection and sequelae: report of an external consultants' meeting**. *Atlanta, GA: Centers for Disease Control and Prevention* 1999.

[10] ACS: **Cancer Facts and Figures 2011**. *Atlanta: American Cancer Society* 2011.

[11] Koutsky LA, Ault KA, Wheeler CM, Brown DR, Barr E, Alvarez FB, Chiacchierini LM, Jansen KU, of Principle Study Investigators P: **A controlled trial of a human papillomavirus type 16 vaccine**. *N Engl J Med* 2002, **347**(21):1645–51.

[12] Liao JB: **Viruses and human cancer**. *Yale J Biol Med* 2006, **79**(3-4):115–22.

[13] Engels EA, Biggar RJ, Hall HI, Cross H, Crutchfield A, Finch JL, Grigg R, Hylton T, Pawlish KS, McNeel TS, Goedert JJ: **Cancer risk in people infected with human immunodeficiency virus in the United States**. *Int J Cancer* 2008, **123**:187–94.

[14] Grulich AE, van Leeuwen MT, Falster MO, Vajdic CM: **Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis**. *Lancet* 2007, **370**(9581):59–67.

[15] Balansky R, D'Agostini F, Micale RT, Maestra SL, Steele VE, Flora SD: **Dose-related cytogenetic damage in pulmonary alveolar macrophages from mice exposed to cigarette smoke early in life**. *Arch Toxicol* 2011.

[16] Higuchi A, Ito K, Dogru M, Kitamura M, Mitani F, Kawakita T, Ogawa Y, Tsubota K: **Corneal damage and lacrimal gland dysfunction in a smoking rat model**. *Free Radic Biol Med* 2011, **51**(12):2210–6.

[17] Dalla-Favera R, Bregni M, Erikson J, Patterson D, Gallo RC, Croce CM: **Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells**. *Proc Natl Acad Sci USA* 1982, **79**(24):7824–7.

[18] Tsujimoto Y, Yunis J, Onorato-Showe L, Erikson J, Nowell PC, Croce CM: **Molecular cloning of the chromosomal breakpoint of B-cell lymphomas and leukemias with the t(11;14) chromosome translocation**. *Science* 1984, **224**(4656):1403–6.

[19] Tsujimoto Y, Cossman J, Jaffe E, Croce CM: **Involvement of the bcl-2 gene in human follicular lymphoma**. *Science* 1985, **228**(4706):1440–3.

[20] Capon DJ, Chen EY, Levinson AD, Seeburg PH, Goeddel DV: **Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue**. *Nature* 1983, **302**(5903):33–7.

[21] McCoy MS, Toole JJ, Cunningham JM, Chang EH, Lowy DR, Weinberg RA: **Characterization of a human colon/lung carcinoma oncogene**. *Nature* 1983, **302**(5903):79–81.

[22] Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JWC, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA: **Mutations of the BRAF gene in human cancer**. *Nature* 2002, **417**(6892):949–54.

[23] Croce CM, Thierfelder W, Erikson J, Nishikura K, Finan J, Lenoir GM, Nowell PC: **Transcriptional activation of an unrearranged and untranslocated c-myc oncogene by translocation of a C lambda locus in Burkitt**. *Proc Natl Acad Sci USA* 1983, **80**(22):6922–6.

[24] Erikson J, Nishikura K, ar Rushdi A, Finan J, Emanuel B, Lenoir G, Nowell PC, Croce CM: **Translocation of an immunoglobulin kappa locus to a region 3' of an unrearranged c-myc oncogene enhances c-myc transcription**. *Proc Natl Acad Sci USA* 1983, **80**(24):7581–5.

[25] Zattara-Cannoni H, Gambarelli D, Lena G, Dufour H, Choux M, Grisoli F, Vagner-Capodano AM: **Are juvenile pilocytic astrocytomas benign tumors? A cytogenetic study in 24 cases**. *Cancer Genet Cytogenet* 1998, **104**(2):157–60.

[26] Jones DTW, Ichimura K, Liu L, Pearson DM, Plant K, Collins VP: **Genomic analysis of pilocytic astrocytomas at 0.97 Mb resolution shows an increasing tendency**

toward chromosomal copy number change with age. *J Neuropathol Exp Neurol* 2006, **65**(11):1049–58.

[27] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer**. *Science* 2005, **310**(5748):644–8.

[28] Heldin CH, Westermark B: **Mechanism of action and in vivo role of platelet-derived growth factor**. *Physiol Rev* 1999, **79**(4):1283–316.

[29] Croce CM: **Oncogenes and cancer**. *N Engl J Med* 2008, **358**(5):502–11, [[http://www.nejm.org/doi/full/10.1056/NEJMra072367]].

[30] Heinrich MC, Blanke CD, Druker BJ, Corless CL: **Inhibition of KIT tyrosine kinase activity: a novel molecular approach to the treatment of KIT-positive malignancies**. *J Clin Oncol* 2002, **20**(6):1692–703.

[31] Talpaz M, Shah NP, Kantarjian H, Donato N, Nicoll J, Paquette R, Cortes J, O'Brien S, Nicaise C, Bleickardt E, Blackwood-Chirchir MA, Iyer V, Chen TT, Huang F, Decillis AP, Sawyers CL: **Dasatinib in imatinib-resistant Philadelphia chromosome-positive leukemias**. *N Engl J Med* 2006, **354**(24):2531–41.

[32] Sordella R, Bell DW, Haber DA, Settleman J: **Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways**. *Science* 2004, **305**(5687):1163–7.

[33] Harris H: **The analysis of malignancy by cell fusion: the position in 1988**. *Cancer Research* 1988, **48**(12):3302–6.

[34] Knudson AG: **Mutation and cancer: statistical study of retinoblastoma**. *Proc Natl Acad Sci USA* 1971, **68**(4):820–3.

[35] Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, Dryja TP: **A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma**. *Nature* 1986, **323**(6089):643–6.

[36] Sparkes RS, Murphree AL, Lingua RW, Sparkes MC, Field LL, Funderburk SJ, Benedict WF: **Gene for hereditary retinoblastoma assigned to human chromosome 13 by linkage to esterase D**. *Science* 1983, **219**(4587):971–3.

[37] Benedict WF, Murphree AL, Banerjee A, Spina CA, Sparkes MC, Sparkes RS: **Patient with 13 chromosome deletion: evidence that the retinoblastoma gene is a recessive cancer gene**. *Science* 1983, **219**(4587):973–5.

[38] Levy DB, Smith KJ, Beazer-Barclay Y, Hamilton SR, Vogelstein B, Kinzler KW: **Inactivation of both APC alleles in human and mouse tumors**. *Cancer Research* 1994, **54**(22):5953–8.

[39] Smith SA, Easton DF, Evans DG, Ponder BA: **Allele losses in the region 17q12-21 in familial breast and ovarian cancer involve the wild-type chromosome**. *Nat Genet* 1992, **2**(2):128–31.

[40] Gudmundsson J, Johannesdottir G, Bergthorsson JT, Arason A, Ingvarsson S, Egilsson V, Barkardottir RB: **Different tumor types from BRCA2 carriers show wild-type chromosome deletions on 13q12-q13**. *Cancer Research* 1995, **55**(21):4830–2.

[41] Bellacosa A, Godwin AK, Peri S, Devarajan K, Caretti E, Vanderveer L, Bove B, Slater C, Zhou Y, Daly M, Howard S, Campbell KS, Nicolas E, Yeung AT, Clapper ML, Crowell JA, Lynch HT, Ross E, Kopelovich L, Knudson AG: **Altered gene expression in morphologically normal epithelial cells from heterozygous carriers of BRCA1 or BRCA2 mutations**. *Cancer Prevention Research* 2010, **3**:48–61.

[42] Burga LN, Tung NM, Troyan SL, Bostina M, Konstantinopoulos PA, Fountzilas H, Spentzos D, Miron A, Yassin YA, Lee BT, Wulf GM: **Altered proliferation and differenti-**

ation properties of primary mammary epithelial cells from BRCA1 mutation carriers. *Cancer Research* 2009, **69**(4):1273–8.

[43] Proia TA, Keller PJ, Gupta PB, Klebba I, Jones AD, Sedic M, Gilmore H, Tung N, Naber SP, Schnitt S, Lander ES, Kuperwasser C: **Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate**. *Cell Stem Cell* 2011, **8**(2):149–63.

[44] Berger AH, Knudson AG, Pandolfi PP: **A continuum model for tumour suppression**. *Nature* 2011, **476**(7359):163–9.

[45] Weinberg RA: **The retinoblastoma protein and cell cycle control**. *Cell* 1995, **81**(3):323–30.

[46] Oren M, Rotter V: **Introduction: p53–the first twenty years**. *Cell Mol Life Sci* 1999, **55**:9–11.

[47] German J, Sanz MM, Ciocci S, Ye TZ, Ellis NA: **Syndrome-causing mutations of the BLM gene in persons in the Bloom's Syndrome Registry**. *Hum Mutat* 2007, **28**(8):743–53.

[48] Amor-Guéret M: **Bloom syndrome, genomic instability and cancer: the SOS-like hypothesis**. *Cancer Lett* 2006, **236**:1–12.

[49] Kadouri L, Hubert A, Rotenberg Y, Hamburger T, Sagi M, Nechushtan C, Abeliovich D, Peretz T: **Cancer risks in carriers of the BRCA1/2 Ashkenazi founder mutations**. *J Med Genet* 2007, **44**(7):467–71.

[50] Thompson D, Easton DF, Consortium BCL: **Cancer Incidence in BRCA1 mutation carriers**. *J Natl Cancer Inst* 2002, **94**(18):1358–65.

[51] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nat Rev Cancer* 2004, **4**(3):177–83.

[52] Chen JM, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP: **Genomic rearrangements in inherited disease and cancer**. *Semin Cancer Biol* 2010, **20**(4):222–33.

[53] De S, Michor F: **DNA secondary structures and epigenetic determinants of cancer genome evolution**. *Nat Struct Mol Biol* 2011, **18**(8):950–5.

[54] Stallings RL: **Are chromosomal imbalances important in cancer?** *Trends Genet* 2007, **23**(6):278–83.

[55] Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL: **Genetic alterations during colorectal-tumor development**. *N Engl J Med* 1988, **319**(9):525–32.

[56] Greenman CD, Pleasance ED, Newman S, Yang F, Fu B, Nik-Zainal S, Jones D, Lau KW, Carter N, Edwards PAW, Futreal PA, Stratton MR, Campbell PJ: **Estimation of rearrangement phylogeny for cancer genomes**. *Genome Research* 2011.

[57] Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, Mclaren S, Lin ML, Mcbride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, Futreal PA, Campbell PJ: **Massive genomic rearrangement acquired in a single catastrophic event during cancer development**. *Cell* 2011, **144**:27–40.

[58] Hanahan D, Weinberg R: **The hallmarks of cancer.** *Cell* 2000, **100**:57–70.

[59] Sherr CJ: **Principles of tumor suppression**. *Cell* 2004, **116**(2):235–46, [[http://www.sciencedirect.com/science/article/pii/S0092867403010754]].

[60] Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Research* 2011, **39**(17):e118, [[http://nar.oxfordjournals.org/content/39/17/e118.long]].

[61] Nowell PC: **The minute chromosome (Phl) in chronic granulocytic leukemia**. *Blut* 1962, **8**:65–6.

[62] Rowley JD: **Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining**. *Nature* 1973, **243**(5405):290–3.

[63] Chissoe SL, Bodenteich A, Wang YF, Wang YP, Burian D, Clifton SW, Crabtree J, Freeman A, Iyer K, Jian L: **Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation**. *Genomics* 1995, **27**:67–82.

[64] Shaulian E, Karin M: **AP-1 in cell proliferation and survival**. *Oncogene* 2001, **20**(19):2390–400.

[65] Shaulian E, Karin M: **AP-1 as a regulator of cell life and death**. *Nat Cell Biol* 2002, **4**(5):E131–6.

[66] Felix M: **Cancer cytogenetics update 2005**. *tlas of Genetics and Cytogenetics in Oncology and Haematology* 2005.

[67] Erikson J, Martinis J, Croce CM: **Assignment of the genes for human lambda immunoglobulin chains to chromosome 22**. *Nature* 1981, **294**(5837):173–5.

[68] Chou WC, Dang CV: **Acute promyelocytic leukemia: recent advances in therapy and molecular basis of response to arsenic therapies**. *Curr Opin Hematol* 2005, **12**:1–6.

[69] Hunger SP: **Chromosomal translocations involving the E2A gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis**. *Blood* 1996, **87**(4):1211–24.

[70] Rotman G, Shiloh Y: **ATM: from gene to function**. *Hum Mol Genet* 1998, **7**(10):1555–63.

[71] Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirieac LR, Cho J, Fujii Y, Gazdar AF, Giordano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, Mardis ER, McPherson JD, Minna JD, Morgan MB, Nadel M, Orringer MB, Osborne JR, Ozenberger B, Ramos AH, Robinson J, Roth JA, Rusch V, Sasaki H, Shepherd F, Sougnez C, Spitz MR, Tsao MS, Twomey D, Verhaak RGW, Weinstock GM, Wheeler DA, Winckler W, Yoshizawa A, Yu S, Zakowski MF, Zhang Q, Beer DG, Wistuba II, Watson MA, Garraway LA, Ladanyi M, Travis WD, Pao W, Rubin MA, Gabriel SB, Gibbs RA, Varmus HE, Wilson RK, Lander ES, Meyerson M: **Characterizing the cancer genome in lung adenocarcinoma**. *Nature* 2007, **450**(7171):893–8.

[72] Lahortiga I, Keersmaecker KD, Vlierberghe PV, Graux C, Cauwelier B, Lambert F, Mentens N, Beverloo HB, Pieters R, Speleman F, Odero MD, Bauters M, Froyen G, Marynen P, Vandenberghe P, Wlodarska I, Meijerink JPP, Cools J: **Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia**. *Nat Genet* 2007, **39**(5):593–5.

[73] Network CGAR: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061–8.

[74] Hastings PJ, Lupski JR, Rosenberg SM, Ira G: **Mechanisms of change in gene copy number**. *Nat Rev Genet* 2009, **10**(8):551–64.

[75] Lieber MR: **The mechanism of human nonhomologous DNA end joining**. *J Biol Chem* 2008, **283**:1–5.

[76] De S: **Somatic mosaicism in healthy human tissues**. *Trends Genet* 2011, **27**(6):217–23.

[77] Kuwabara T, Hsieh J, Muotri A, Yeo G, Warashina M, Lie DC, Moore L, Nakashima K, Asashima M, Gage FH: **Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis**. *Nat Neurosci* 2009, **12**(9):1097–105.

[78] Li Y, Benezra R: **Identification of a human mitotic checkpoint gene: hsMAD2**. *Science* 1996, **274**(5285):246–8.

[79] Milner J, Ponder B, Hughes-Davies L, Seltmann M, Kouzarides T: **Transcriptional activation functions in BRCA2**. *Nature* 1997, **386**(6627):772–3.

[80] Yarden RI, Pardo-Reoyo S, Sgagias M, Cowan KH, Brody LC: **BRCA1 regulates the G2/M checkpoint by activating Chk1 kinase upon DNA damage**. *Nat Genet* 2002, **30**(3):285–9.

[81] Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger N, Remke M, Shih D, Northcott PA, Pfaff E, Tica J, Wang Q, Massimi L, Witt H, Bender S, Pleier S, Cin H, Hawkins C, Beck C, von Deimling A, Hans V, Brors B, Eils R, Scheurlen W, Blake J, Benes V, Kulozik AE, Witt O, Martin D, Zhang C, Porat R, Merino DM, Wasserman J, Jabado N, Fontebasso A, Bullinger L, Rücker FG, Döhner K, Döhner H, Koster J, Molenaar JJ, Versteeg R, Kool M, Tabori U, Malkin D, Korshunov A, Taylor MD, Lichter P, Pfister SM, Korbel JO: **Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations**. *Cell* 2012, **148**(1-2):59–71.

[82] Painter TS: **The spermatogenesis of man**. *Anat Res* 1922, **23**:129.

[83] Painter TS: **Studies in mammalian spermatogenesis II. The spermatogenesis of man.** *J Exp Zoology* 1923, **37**:291–336.

[84] H TJ, A L: **The chromosome number of man.** *Hereditas* 1956, **42**:1–6.

[85] M K: **From 48 to 46: cytological technique, preconception and the counting of the human chromosomes**. *Bull Hist Med* 1974, **48**:465–502.

[86] FORD CE, HAMERTON JL: **The chromosomes of man**. *Nature* 1956, **178**(4541):1020–3.

[87] Caspersson T, Zech L, Johansson C: **Analysis of human metaphase chromosome set by aid of DNA-binding fluorescent agents**. *Exp Cell Res* 1970, **62**(2):490–2.

[88] Watt JL SG: **Lymphocyte culture for chromosome analysis. In:RooneyDE, CzepulkowskiBH, eds.** *Human cytogenetics:a practical approach* 1986, **1**:39–55.

[89] Mitelman F JB, (Eds) MF: **Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer**. *http://cgap.nci.nih.gov/Chromosomes/Mitelman*.

[90] Baudis M, Cleary ML: **Progenetix.net: an online repository for molecular cytogenetic aberration data**. *Bioinformatics* 2001, **17**(12):1228–9.

[91] Veldman T, Vignon C, Schröck E, Rowley JD, Ried T: **Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping**. *Nat Genet* 1997, **15**(4):406–10.

[92] Cremer T, Lichter P, Borden J, Ward DC, Manuelidis L: **Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes**. *Hum Genet* 1988, **80**(3):235–46.

[93] Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: **Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia**. *Science* 1985, **230**(4732):1350–4.

[94] Lui WO, Kytölä S, Anfalk L, Larsson C, Farnebo LO: **Balanced translocation (3;7)(p25;q34): another mechanism of tumorigenesis in follicular thyroid carcinoma?** *Cancer Genet Cytogenet* 2000, **119**(2):109–12.

[95] Schröck E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, Garini Y, Ried T: **Multicolor spectral karyotyping of human chromosomes**. *Science* 1996, **273**(5274):494–7.

[96] Speicher MR, Ballard SG, Ward DC: **Karyotyping human chromosomes by combinatorial multi-fluor FISH**. *Nat Genet* 1996, **12**(4):368–75.

[97] Telenius H, Pelmear AH, Tunnacliffe A, Carter NP, Behmel A, Ferguson-Smith MA, Nordenskjöld M, Pfragner R, Ponder BA: **Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes**. *Genes Chromosomes Cancer* 1992, **4**(3):257–63.

[98] Meltzer PS, Guan XY, Burgess A, Trent JM: **Rapid generation of region specific probes by chromosome microdissection and their application**. *Nat Genet* 1992, **1**:24–8.

[99] Cram LS, Gray JW, Carter NP: **Cytometry and genetics**. *Cytometry A* 2004, **58**:33–6.

[100] **http://www.ncbi.nlm.nih.gov/sky/ccap**$_h$*elper.cgi?tsc* = 4.

[101] Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P: **Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe**. *Hum Genet* 1993, **90**(6):584–9.

[102] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors**. *Science* 1992, **258**(5083):818–21.

[103] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays**. *Nat Genet* 1999, **23**:41–6.

[104] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays**. *Nat Genet* 1998, **20**(2):207–11.

[105] Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray

JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of DNA copy number**. *Nat Genet* 2001, **29**(3):263–4.

[106] Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IPM, Carter NP: **DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones**. *Genes Chromosomes Cancer* 2003, **36**(4):361–74.

[107] Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nat Rev Genet* 2006, **7**(2):85–97, [[http://www.nature.com/nrg/journal/v7/n2/full/nrg1767.html]].

[108] Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B: **High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides**. *J Clin Pathol* 2004, **57**(6):644–6.

[109] Zhao X, Li C, Paez JG, Chin K, Jänne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M: **An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays**. *Cancer Research* 2004, **64**(9):3060–71.

[110] Slater HR, Bailey DK, Ren H, Cao M, Bell K, Nasioulas S, Henke R, Choo KHA, Kennedy GC: **High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs**. *Am J Hum Genet* 2005, **77**(5):709–26.

[111] Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH: **Whole genome DNA copy number changes identified by high density oligonucleotide arrays**. *Hum Genomics* 2004, **1**(4):287–99.

[112] Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, Futreal PA, Weber B, Shapero MH, Wooster R: **High-resolution analysis of DNA copy number using oligonucleotide microarrays**. *Genome Research* 2004, **14**(2):287–95.

[113] Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, Villapakkam A, Stanton SE, Larsson C, Hudson TJ, Johnson BE, Lander ES, Meyerson M: **Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays**. *Nat Biotechnol* 2000, **18**(9):1001–5.

[114] Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ: **Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays**. *Genome Research* 2000, **10**(8):1126–37.

[115] Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Henry KTM, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M: **The landscape of somatic copy-number alteration across human cancers**. *Nature* 2010, **463**(7283):899–905.

[116] Baudis M: **Online database and bioinformatics toolbox to support data mining in cancer cytogenetics**. *BioTechniques* 2006, **40**(3):269–70, 272.

[117] Baudis M: **Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data**. *BMC Cancer* 2007, **7**:226.

[118] Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Børresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors**. *Proc Natl Acad Sci USA* 2002, **99**(20):12963–8.

[119] Wolf M, Mousses S, Hautaniemi S, Karhu R, Huusko P, Allinen M, Elkahloun A, Monni O, Chen Y, Kallioniemi A, Kallioniemi OP: **High-resolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression**. *Neoplasia* 2004, **6**(3):240–7.

[120] Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, Martin ES, Yang Z, Ji H, Chin L, Depinho RA: **High-resolution genomic profiles of human lung cancer**. *Proc Natl Acad Sci USA* 2005, **102**(27):9625–30.

[121] McArdle L, McDermott M, Purcell R, Grehan D, O'Meara A, Breatnach F, Catchpoole D, Culhane AC, Jeffery I, Gallagher WM, Stallings RL: **Oligonucleotide microarray analysis of gene expression in neuroblastoma displaying loss of chromosome 11q**. *Carcinogenesis* 2004, **25**(9):1599–609.

[122] Wang Q, Diskin S, Rappaport E, Attiyeh E, Mosse Y, Shue D, Seiser E, Jagannathan J, Shusterman S, Bansal M, Khazi D, Winter C, Okawa E, Grant G, Cnaan A, Zhao H, Cheung NK, Gerald W, London W, Matthay KK, Brodeur GM, Maris JM: **Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number**. *Cancer Research* 2006, **66**(12):6050–62.

[123] Janoueix-Lerosey I, Novikov E, Monteiro M, Gruel N, Schleiermacher G, Loriod B, Nguyen C, Delattre O: **Gene expression profiling of 1p35-36 genes in neuroblastoma**. *Oncogene* 2004, **23**(35):5912–22.

[124] Upender MB, Habermann JK, McShane LM, Korn EL, Barrett JC, Difilippantonio MJ, Ried T: **Chromosome transfer induced aneuploidy results in complex dysregulation of the cellular transcriptome in immortalized and cancer cells**. *Cancer Research* 2004, **64**(19):6941–9.

[125] Cox C, Bignell G, Greenman C, Stabenau A, Warren W, Stephens P, Davies H, Watt S, Teague J, Edkins S, Birney E, Easton DF, Wooster R, Futreal PA, Stratton MR: **A survey of homozygous deletions in human cancer genomes**. *Proc Natl Acad Sci USA* 2005, **102**(12):4542–7.

[126] Kwabi-Addo B, Giri D, Schmidt K, Podsypanina K, Parsons R, Greenberg N, Ittmann M: **Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression**. *Proc Natl Acad Sci USA* 2001, **98**(20):11563–8.

[127] Lee Y, Miller HL, Russell HR, Boyd K, Curran T, McKinnon PJ: **Patched2 modulates tumorigenesis in patched1 heterozygous mice**. *Cancer Research* 2006, **66**(14):6964–71.

[128] David G, Dannenberg JH, Simpson N, Finnerty PM, Miao L, Turner GM, Ding Z, Carrasco R, Depinho RA: **Haploinsufficiency of the mSds3 chromatin regulator promotes chromosomal instability and cancer only upon complete neutralization of p53**. *Oncogene* 2006, **25**(56):7354–60.

[129] Snijders AM, Schmidt BL, Fridlyand J, Dekker N, Pinkel D, Jordan RCK, Albertson DG: **Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma**. *Oncogene* 2005, **24**(26):4232–42.

[130] Albertson DG, Ylstra B, Segraves R, Collins C, Dairkee SH, Kowbel D, Kuo WL, Gray JW, Pinkel D: **Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene**. *Nat Genet* 2000, **25**(2):144–6.

[131] Ashton-Rickardt P, Wyllie A, Bird C, Dunlop M, Steel C, Morris R, Piris J, Romanowski P, Wood R, White R, al et: **MCC, a candidate familial polyposis gene in 5q.21, shows frequent allele loss in colorectal and lung cancer.** *Oncogene* 1991, **6**(10):1881–6.

[132] Bergamaschi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Børresen-Dale AL, Pollack JR: **Distinct patterns of DNA copy number**

alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosom. Cancer* 2006, **45**(11):1033–1040.

[133] Chapiro E, Leporrier N, Radford-Weiss I, Bastard C, Mossafa H, Leroux D, Tigaud I, Braekeleer MD, Terré C, Brizard F: **Gain of the short arm of chromosome 2 (2p) is a frequent recurring chromosome aberration in untreated chronic lymphocytic leukemia (CLL) at advanced stages**. *Leukemia Research* 2010, **34**:63–68.

[134] Attiyeh EF, London WB, Mossé YP, Wang Q, Winter C, Khazi D, McGrady PW, Seeger RC, Look AT, Shimada H, Brodeur GM, Cohn SL, Matthay KK, Maris JM, Group CO: **Chromosome 1p and 11q deletions and outcome in neuroblastoma**. *N Engl J Med* 2005, **353**(21):2243–53.

[135] Spitz R, Hero B, Ernestus K, Berthold F: **Deletions in chromosome arms 3p and 11q are new prognostic markers in localized and 4s neuroblastoma**. *Clin Cancer Res* 2003, **9**:52–8.

[136] astowska M, Cotterill S, Bown N, Cullinane C, Variend S, Lunec J, Strachan T, Pearson ADJ, Jackson MS: **Breakpoint position on 17q identifies the most aggressive neuroblastoma tumors**. *Genes Chromosomes Cancer* 2002, **34**(4):428–36.

[137] Al-Kuraya K, Schraml P, Torhorst J, Tapia C, Zaharieva B, Novotny H, Spichtin H, Maurer R, Mirlacher M, Köchli O, Zuber M, Dieterich H, Mross F, Wilber K, Simon R, Sauter G: **Prognostic relevance of gene amplifications and coamplifications in breast cancer**. *Cancer Research* 2004, **64**(23):8534–40.

[138] Rennstam K, Ahlstedt-Soini M, Baldetorp B, Bendahl PO, Borg A, Karhu R, Tanner M, Tirkkonen M, Isola J: **Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization**. *Cancer Research* 2003, **63**(24):8861–8.

[139] Rajagopalan H, Nowak MA, Vogelstein B, Lengauer C: **The significance of unstable chromosomes in colorectal cancer**. *Nat Rev Cancer* 2003, **3**(9):695–701.

[140] Mottolese M, Nádasi EA, Botti C, Cianciulli AM, Merola R, Buglioni S, Benevolo M, Giannarelli D, Marandino F, Donnorso RP, Venturo I, Natali PG: **Phenotypic changes of p53, HER2, and FAS system in multiple normal tissues surrounding breast cancer**. *J Cell Physiol* 2005, **204**:106–12.

[141] Roh HJ, Shin DM, Lee JS, Ro JY, Tainsky MA, Hong WK, Hittelman WN: **Visualization of the timing of gene amplification during multistep head and neck tumorigenesis**. *Cancer Research* 2000, **60**(22):6496–502.

[142] Bastian BC, Kashani-Sabet M, Hamm H, Godfrey T, Moore DH, Bröcker EB, LeBoit PE, Pinkel D: **Gene amplifications characterize acral melanoma and permit the detection of occult tumor cells in the surrounding skin**. *Cancer Research* 2000, **60**(7):1968–73.

[143] Takeuchi I, Tagawa H, Tsujikawa A, Nakagawa M, Katayama-Suguro M, Guo Y, Seto M: **The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of B-cell lymphomas and genetic regions involved in lymphomagenesis**. *Haematologica* 2009, **94**:61–69.

[144] Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, Knuutila S: **DNA copy number amplification profiling of human neoplasms**. *Oncogene* 2006, **25**(55):7324–7332.

[145] Mass RD, PREss MF, Anderson S, Cobleigh MA, Vogel CL, Dybdal N, Leiberman G, SLAMON DJ: **Evaluation of clinical outcomes according to HER2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab**. *Clin Breast Cancer* 2005, **6**(3):240–6.

[146] Wang TL, Diaz LA, Romans K, Bardelli A, Saha S, Galizia G, Choti M, Donehower R, Parmigiani G, Shih IM, Iacobuzio-Donahue C, Kinzler KW, Vogelstein B, Lengauer C, Velculescu VE: **Digital karyotyping identifies thymidylate synthase amplification as a mechanism of resistance to 5-fluorouracil in metastatic colorectal cancer patients**. *Proc Natl Acad Sci USA* 2004, **101**(9):3089–94.

[147] Gorlick R, Goker E, Trippett T, Waltham M, Banerjee D, Bertino JR: **Intrinsic and acquired resistance to methotrexate in acute leukemia**. *N Engl J Med* 1996, **335**(14):1041–8.

[148] Klijn C, Bot J, Adams DJ, Reinders M, Wessels L, Jonkers J: **Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach**. *PLoS Comput Biol* 2010, **6**:e1000631.

[149] Bredel M, Scholtens DM, Harsh GR, Bredel C, Chandler JP, Renfrow JJ, Yadav AK, Vogel H, Scheck AC, Tibshirani R, Sikic BI: **A network model of a cooperative genetic landscape in brain tumors**. *JAMA: The Journal of the American Medical Association* 2009, **302**(3):261–75.

[150] Kumar N, Rehrauer H, Cai H, Baudis M: **CDCOCA: a statistical method to define complexity dependence of co-occuring chromosomal aberrations**. *BMC Med Genomics* 2011, **4**:21.

[151] Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control**. *Nat Med* 2004, **10**(8):789–99, [[http://www.nature.com/nm/journal/v10/n8/full/nm1087.html]].

[152] Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VEH, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, Sauvage FJD, Faham M, Seshagiri S: **Diverse somatic mutation patterns and pathway alterations in human cancers**. *Nature* 2010, **466**(7308):869–873.

[153] Alloza E, Al-Shahrour F, Cigudosa JC, Dopazo J: **A large scale survey reveals that chromosomal copy-number alterations significantly affect gene modules involved in cancer initiation and progression**. *BMC Med Genomics* 2011, **4**:37.

[154] Lee JM: **Genomic Gene Clustering Analysis of Pathways in Eukaryotes**. *Genome Research* 2003, **13**(5):875–882.

[155] Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M: **Distance-based clustering of CGH data**. *Bioinformatics* 2006, **22**(16):1971–8.

[156] Ferreira BI, Garcia JF, Suela J, Mollejo M, Camacho FI, Carro A, Montes S, Piris MA, Cigudosa JC: **Comparative genome profiling across subtypes of low-grade B-cell lymphoma identifies type-specific and common aberrations that target genes with a role in B-cell neoplasia**. *Haematologica* 2008, **93**(5):670–679.

[157] Liu J, Ranka S, Kahveci T: **Markers improve clustering of CGH data**. *Bioinformatics* 2007, **23**(4):450–7.

[158] Wieringen WNV, Wiel MAVD, Ylstra B: **Weighted clustering of called array CGH data**. *Biostatistics* 2008, **9**(3):484–500, [[http://biostatistics.oxfordjournals.org/content/9/3/484.long]].

[159] Kumar N, Cai H, v Mering C, Baudis M: **Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated molecular-cytogenetic data**. *Submitted*.

[160] Joshi-Tope G: **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Research* 2004, **33**(Database issue):D428–D432.

[161] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Research* 2010, **38**(Database):D355–D360.

[162] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic Acids Research* 2009, **37**(Database):D674–D679.

[163] Curtis RK, Orešič M, Vidal-Puig A: **Pathways to the analysis of microarray data**. *Trends in Biotechnology* 2005, **23**(8):429–435.

[164] Cai H, Kumar N, Baudis M: **arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies**. *PLoS ONE* 2012, **7**(5):e36944,